Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jernej Vičič

Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike

DOKTORSKA DISERTACIJA

Mentor: prof. dr. Igor Kononenko
Somentor: doc. dr. Tomaž Erjavec

Ljubljana, 2012

Zahvala

Iskreno se zahvaljujem mentorju prof. dr. Igorju Kononenku in somentorju doc. dr. Tomažu Erjavcu za strokovno pomoč in usmeritve pri izdelavi doktorske disertacije.

Zahvaljujem se dr. Karin Marc Bratina za pomoč pri razvozlavanju jezikoslovnih ugank ter za kritičen pregled in tehtne predloge pri urejanju doktorske disertacije.

Posebej bi se rad zahvalil svoji družini, ki me je podpirala na dolgotrajni poti do tega izdelka.

Hvala!

I Z J A V A   O   A V T O R S T V U

doktorske disertacije

Spodaj podpisani/-a Jernej Vičič,

z vpisno številko 24930372,

sem avtor/-ica doktorske disertacije z naslovom

Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike
S svojim podpisom zagotavljam, da:

- sem doktorsko disertacijo izdelal/-a samostojno pod vodstvom mentorja
  prof. dr. Igorja Kononenka

  in somentorstvom

  doc. dr. Tomaža Erjavca
- so elektronska oblika doktorske disertacije, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko doktorske disertacije
- in soglašam z javno objavo elektronske oblike doktorske disertacije v zbirki "Dela FRI".

V Ljubljani, dne ________________ Podpis avtorja/-ice: ________________

Mojim.

Kazalo

Seznam uporabljenih kratic in simbolov

BLEU Bilingual Evaluation Understudy, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo uspešnih prevodov

CBMT Corpus Based Machine Translation, strojno prevajanje na osnovi korpusov

EBMT Example Based Machine Translation, strojno prevajanje na osnovi primerov

FAMT Fully Automatic Machine Translation, popolnoma samodejno strojno prevajanje

GNU GNU's Not Unix!, projekt za izdelavo operacijskega sistema GNU

GPL GNU General Public License, prosta licenca za programsko opremo in ostalo

HMT Hybrid Machine Translation, hibridno (mešano) strojno prevajanje

LDC Linguistic Data Consortium, odprti konzorcij članov, ki se ukvarjajo z jezikovnimi tehnologijami

LGPL GNU Lesser General Public License, prosta licenca za programsko opremo in ostalo z omiljenimi omejitvami

METEOR Metric for Evaluation of Translation with Explicit ORdering, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo uspešnih prevodov

MSD Morphosyntactic description, oblikoskladenjska oznaka

MT Machine Translation, strojno prevajanje

NIST National Institute of Standards and Technology, inštitut za standarde in tehnologijo

PBMT Phrase Based Machine Translation, strojno prevajanje na osnovi fraz

PoS Part of Speech, označba besedne vrste

RBMT Rule Based Machine Translation, strojno prevajanje na osnovi pravil

SMT Statistical Machine Translation, statistično strojno prevajanje

WER Word Error Rate, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo napake

WFST Weighed Finite State Transducers, uteženi končni avtomati z izhodom

WRR Word Recognition Rate, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo uspešnih prevodov

XML Extensible Markup Language, jezik za označevanje

Povzetek

Pričujoče delo predstavlja pregled strojnega prevajanja naravnih jezikov, osredotoča se predvsem na sisteme in metode za prevajanje sorodnih naravnih jezikov. Večina predstavljenih sistemov sodi v skupino strojnega prevajanja na osnovi pravil plitkega prenosa, ki so najprimernejši za postavitev sistemov za strojno prevajanje sorodnih jezikov. Največja težava sistemov, ki temeljijo na pravilih, je dolgotrajna in draga ročna izdelava slovarjev ter prevajalnih pravil v primeru klasičnega pristopa h gradnji prevajalnih sistemov na osnovi pravil. Delo ponuja pregled zbirke izbranih in na novo zasnovanih metod samodejne izdelave gradiv za postavitev prevajalnih sistemov na osnovi pravil. Metode so bile preizkušene na študiji primera: postavitev popolnoma delujočega prevajalnega sistema za sorodne jezike. Postavljeni so bili štirje sistemi: slovenščina-srbščina, slovenščina-češčina, slovenščina-angleščina in slovenščina-estonščina. Poleg same kakovosti prevodov je bila ocenjena tudi hitrost postavitve novega prevajalnega sistema.

V disertaciji je predstavljena metoda, ki razširja osnovno metodo za prevajanje s pomočjo dreves izpeljav za jezike z omejeno podporo jezikovnih tehnologij. V učni fazi je za izvorni jezik namesto drevesnice uporabljen le poravnani korpus.

V disertaciji je opisana metoda za samodejno izdelavo oblikoskladenjskih slovarjev, ki vključuje samodejno označevanje paradigem, njihovo samodejno luščenje za visoko pregibne jezike in izdelavo pripadajočih leksikonov ter samodejno izdelavo dvojezičnih prevajalnih slovarjev.

V disertaciji je predstavljena metoda za uporabo, izbiro in ocenjevanje pravil za strukturni prenos. Opisane metode za samodejno gradnjo pravil strukturnega prenosa pogosto izdelajo veliko množico pravil, ki med sabo tekmujejo (mogoče jih je uporabiti za iste dele besedila). Najboljša pravila izberemo na podlagi korpusa ciljnega jezika.

Ključne besede: strojno prevajanje, strojno prevajanje sorodnih naravnih jezikov, tehnologije hitrih postavitev sistemov za strojno prevajanje

Abstract

The work presents an overview of the systems and methods for the natural language machine translation. It focuses primarily on systems and methods for the translation of the related languages. Most of the presented systems belong to the Shallow Parse and Transfer Rule-Based Machine Translation paradigm, which is better suited for the implementation of a translation system for related languages. The major problem of the rule-based translation systems is costly manual production of dictionaries and translation rules in the case of a classical approach to building such systems. The work provides an overview over the collection of selected and new methods designed for automatic production of materials for the installation of systems based on translation rules.

Methods were tested on a case study: the implementation of a fully functioning translation system for related languages. The following four systems were used as the basis: Slovenian-Serbian, Slovenian-Czech, Slovenian-English and Slovenian-Estonian. The evaluation process focused on the quality of the translations as well as the estimation of the time needed for the implementation of a new system.

The dissertation presents a method that extends the basic Statistical Machine Translation by Parsing paradigm for languages with limited support of language technologies. The learning phase uses an aligned corpus instead of a full treebank.

The dissertation describes a method for the automatic creation of morphologies, which includes automatic paradigm tagging, automatic paradigm construction for the highly inflected languages and automatic production of bilingual dictionaries.

The dissertation presents a method for the selection and assessment of the rules for the structural transfer. Methods for the automatic construction of structural transfer rules often produce a large set of rules, which compete with each other (it is possible to use multiple rules on the same part of text). The best rules are chosen on the basis of the target language corpus.

Key words: rbmt, machine translation, machine translation of related languages, speeding up the implementation of machine translation systems

Poglavje 1

Uvod

1.1 Motivacija

Sorodnost naravnih jezikov ene tipološke skupine in včasih celo jezikov različnih tipoloških skupin (primer poskusa prevajalnega sistema za jezikovni par češčina-litovščina (Hajič et al., 2003)) omogoča lažje in natančnejše prevajanje ter uporabo enostavnejših metod, ki ne bi bile dovolj dobre za uporabo v prevajalnih sistemih nesorodnih jezikovnih parov. Uporaba preprostejših metod ne pomeni nujno slabše kakovosti prevodov; veliko napak sistemov za prevajanje namreč izvira ravno iz napak v skladenjskem razčlenjevanju (parsing) izvornih povedi. Seštevanje napak v analizi, prenosu in generiranju pri sistemih za strojno prevajanje na osnovi pravil s klasično arhitekturo pogosto prinese slabše rezultate kot uporaba enostavnih metod plitkega razčlenjevanja in prenosa.

Ena od glavnih ovir, ki upočasnjujejo proces razvoja prevajalnih sistemov na osnovi pravil, je obseg človeškega dela, ki je nujno za oblikovanje pravil slovnice in slovarjev. S tem problemom se soočajo tudi sistemi, ki so namenjeni prevajanju sorodnih jezikov. Takšni sistemi navadno uporabljajo poenostavljeno arhitekturo in izkoriščajo podobnost jezikov z uporabo plitke slovnice in pravil prenosa, vendar tudi ta pravila zahtevajo veliko napora.

Pričujoče delo predstavlja sisteme za strojno prevajanje sorodnih jezikov. Prikazan je pregled že izdelanih prevajalnih sistemov za izbrane jezikovne pare. Opisani so osnovne komponente in jezikovna gradiva, ki jih uporablja večina teh sistemov. Poseben poudarek je na opisu komponent in gradiv zbirke orodij ter gradiv za postavitev novih sistemov za strojno prevajanje sorodnih jezikov Apertium (Corbi-Bellot et al., 2005). V tem delu so med drugim predstavljene metode, ki omogočajo samodejno ustvarjanje vseh gradiv, ki so potrebna za postavitev sistema za prevajanje naravnih sorodnih jezikov.

1.2 Obstoječi sistemi strojnega prevajanja sorodnih jezikov

Ruslan (Hajič, 1987) je prvi sistem za strojno prevajanje sorodnih jezikov. Prevajalni par sistema je bil češčina-ruščina. V sistemu je bilo uporabljeno globoko skladenjsko razčlenjevanje (deep syntactic parsing) in prenos. Uporaba je bila omejena na prevajanje uporabniških navodil.

Česílko (Hajič et al., 2000) je sistem za strojno prevajanje sorodnih jezikov, in sicer češčine in slovaščine. Arhitektura osnovne različice je bila enostavna, slovarji z direktnimi prevodi lem ena-na-ena z leksikalnim prenosom in brez dodatnih pravil. V sistemu so bile uporabljene metode, ki so temeljile na dejstvu, da sta si jezika zelo podobna. Kasneje je bil sistem izpopolnjen (Łukasz Dębowski et al., 2002), dodan mu je bil tudi nov jezikovi par, češčina-poljščina.

Osnovna arhitektura sistema je naslednja:

- oblikoskladenjsko označevanje izvornega besedila,
- dvojezični slovarji,
- oblikoskladenjska sinteza v ciljno besedilo.

V novi različici sistema (Hajič et al., 2003) je bil implementiran plitki prenos. Osnovna arhitektura spremenjenega sistema:

- oblikoskladenjska analiza izvornega besedila,
- oblikoskladenjsko razdvoumljanje,
- leksikalni/oblikoskladenjski prenos,
- oblikoskladenjska sinteza v ciljno besedilo.

Natančnost prevodov z metodo WRR (Vogel et al., 2000) je okrog $90\,\%$ za jezikovni par češčina-slovaščina in $71,4\,\%$ za jezikovni par češčina-poljščina.

Guat (Vičič, 2009) je sistem plitkega prenosa, temelječ na ogrodju Apertium. Sistem podpira jezikovni par slovenščina-srbščina. Natančneje je predstavljen v razdelku 2.5.

PONS (Dyvik, 1995) je sistem za delno prevajanje med sorodnimi jeziki (partial translation between related languages). Izdelan je bil za jezikovni par norveščina-švedščina. Zanimiva lastnost sistema je, da ne uporablja oblikoskladenjske analize, slovar izvornega jezika hrani vse besedne oblike. Sistem uporablja delno skladenjsko razčlenjevanje izvornih povedi, ki jih razdeli na kose in manjše enote razčleni.

T4F (Ahrenberg in Holmqvist, 2004) je drugi prevajalni sistem, namenjen skandinavskim jezikom. Kratica imena pomeni tokenizacija, označevanje, prenos, transpozicija in filtriranje (Tokenization, Tagging, Transfer, Transposition and Filtering). Avtorji sistema trdijo, da za strukturno sorodne jezike abstraktno skladenjsko razčlenjevanje ni potrebno oziroma prinaša slabe rezultate.

Prevajalni sistem za Turkijske jezike. Altintas in Cicekli (2002) sta postavila sistem za strojno prevajanje sorodnih turkijskih jezikov, in sicer na podlagi prevajalnega para turščina-krimska tatarščina (Crimean tatar). Avtorji trdijo, da za jezike s skupno zgodovino in podobno kulturo ne potrebujemo semantične analize. Sistem se osredotoča na razlike na oblikoslovni ravni.

Apertium (Corbi-Bellot et al., 2005) je zbirka orodij za postavitev prevajalnih sistemov za sorodne jezike; predstavljen je v razdelku 3.4. Apertium je bil najprej zastavljen kot orodje za postavitev sistemov za strojno prevajanje sorodnih romanskih jezikov; tako so nastali tudi prvi jezikovni pari katalonščina-španščina, španščina-portugalščina in katalonščina-portugalščina.

Prevajalni sistem za Keltske jezike. Scannell (2006) je postavil sistem za strojno prevajanje med irščino (Irish) in škotsko gelščino (Scotish Gaelic). Sistem temelji na ogrodju Apertium. Jezika sta si slovnično sorodna, saj imata skupnega prednika – srednjo irščino (Middle Irish).

1.3 Pregled vsebine

Drugo poglavje prinaša pregled raziskovalnega področja in razlag osnovnih pojmov znanstvenega področja, ki bralcu približajo področje in mu omogočijo nadaljnje branje. Tretje poglavje opisuje eno od možnih razdelitev strojnega prevajanja z opisom posameznih načinov strojnega prevajanja. Poseben poudarek je posvečen prevajanju sorodnih oziroma nesorodnih jezikov v omejenih domenah in ogrodju Apertium, ki je osnovna platforma za večino implementacij metod, opisanih v tem delu. Četrto poglavje podaja oblikoskladenjsko označene slovarje, enojezične in večjezične. Podane so metode za samodejno izdelavo oblikoskladenjsko označenih slovarjev, ki so uporabljeni v strojnih prevajalnih sistemih. Peto poglavje predstavlja pravila prenosa, ki pri strojnih prevajalnih sistemih na osnovi pravil omogočajo opisovanje razlik med jezikoma jezikovnega para. V šestem poglavju je opisana metoda, ki omogoča izdelavo sistema za strojno prevajanje na osnovi dreves izpeljav za manj uporabljene jezike oziroma za jezike, ki nimajo izdelanega skladenjsko označenega dvojezičnega korpusa. V sedmem poglavju so opisane osnove vrednotenja sistemov za strojno prevajanje; predstavljene so uporabljene metrike in metodologije vrednotenja. V zadnjem delu poglavja so podani rezultati vrednotenj posameznih sistemov, zgrajenih na osnovi metod, predstavljenih v četrtem in petem poglavju. Osmo poglavje zaključuje delo z razpravo in s smernicami za nadaljnje delo. V prilogi A so prikazani primeri pravil prenosa in pravil za ujemanje oblikoskladenjskih kategorij bližnjih besed. V prilogi B so zbrani primeri prevodov predstavljenih sistemov.

Poglavje 2

Pregled področja

Pričujoče poglavje uvaja bralca v področje. Predstavljeni so osnovni pojmi, ki so potrebni za lažje razumevanje besedila. Opisani so slovanski jeziki in zlasti njihove posebnosti. Večina metod, opisanih v tem delu, je bila preizkušena prav na sistemih za prevajanje slovanskih jezikov. Nadaljuje se z opisom podobnosti slovanskih jezikov, ki jih izkoriščamo pri izdelavi prevajalnikov, sledi predstavitev uporabljenih učnih gradiv in orodij. Predzadnji razdelek predstavi prevajalni sistem Guat, kjer je implementirana večina predstavljenih metod. Poglavje se zaključi s predstavitvijo in umestitvijo izvirnih prispevkov k znanosti.

2.1 Osnovni pojmi

V nadaljevanju so razloženi osnovni pojmi s področja jezikovnih tehnologij, predvsem s področja strojnega prevajanja naravnih jezikov. Razdelek je namenjen zlasti bralcem, ki so jim jezikovne tehnologije tuje.

2.1.1 Pregibno oblikoslovje

Pregibno oblikoslovje (inflectional morphology), kot ga razlaga Janda (2007), je del slovnice, ki se ukvarja s pregibanjem besed. Zadeva različne pregibne oblike leksemov. V večini jezikov označuje razmerja med osebo, številom, sklonom, spolom, časom in drugimi lastnostmi. Le težko ga umestimo v enotno področje, zato ga uvrščamo med besedoslovje (leksiko) in skladnjo (sintakso) jezika.

2.1.1.1 Besedni razredi

V slovnici je besedni razred jezikoslovna kategorija besed, ki je v splošnem definirana s skladenjskim ali oblikoslovnim obnašanjem besed, ki sodijo v to kategorijo. Skupne jezikovne kategorije vključujejo na primer samostalnik, glagol, pridevnik. V zadnjem času se v literaturi uveljavlja delitev označevanja oblikoskladenjskih kategorij glede na zbirke oznak:

- natančno označevanje (fine-granied PoS tagging) oziroma označevaje oblikoskladenjskih lastnosti besed je natančneje predstavljeno v razdelku 2.1.1.2;
- grobo označevanje (coarse-(PoS) tagsets) za vsako besedo poišče (označi) le njeno glavno lastnost (besedno vrsto), to je PoS.

V nadaljevanju dela se pojem oblikoskladenjskega označevanja nanaša na natančno označevanje (fine-granied PoS tagging), uporabljajo se oblikoskladenjske oznake (MSD – morphosyntactic descriptions) po specifikacijah projekta MULTEXT(-East) (Erjavec, 2010; Dimitrova et al., 1998).

2.1.1.2 Označevanje besednih razredov

Posamezne besede v besedilu razvrstimo v najprimernejše besedne razrede (definirani v prejšnjem razdelku) upoštevajoč definicijo besede in tudi njeno okolico v besedilu (povezava z okoliškimi besedami). Besedni razredi so predstavljeni z ustreznimi oznakami MSD,

Označevanje MSD (morphosyntactic descriptions) je težji problem, kot samo uporaba seznama besed z ustreznimi oznakami, saj lahko besedam pripišemo različne oznake MSD, odvisno od uporabe v besedilu. Največja problema označevanja MSD sta odpravljanje dvoumnosti (disambiguation), to je izbiranje najprimernejše oznake v odvisnosti od konteksta v primeru več možnih oznak, in označevanje neznanih besed.

V literaturi se za oblikoskladenjsko označevanje pojavlja več terminov:

- morfosintaktično označevanje,
- označevanje MSD,
- natančno označevanje (fine-granied PoS tagging).

V slovenskih korpusih so standardne oznake MSD po oblikoskladenjskih specifikacijah projekta MULTEXT(-East) (Erjavec, 2010; Dimitrova et al., 1998), ki temeljijo na delu skupine EAGLES (Calzolari in Monachini, 1996) ter določajo strukturo in vsebino veljavnih oblikoskladenjskih oznak ali MSD-jev (morphosyntactic descriptions). Specifikacije za vsak posamezen jezik opredeljujejo, katere so veljavne oznake in kaj pomenijo. Tako na primer določajo, da je MSD s črkovnim nizom *Sosei* veljaven za označevanje slovenščine in je ekvivalenten naboru naslednjih lastnosti:

- samostalnik,
- vrsta=občno_ime,
- spol=srednji,
- število=ednina,
- sklon=imenovalnik.

Ena od besednih oblik, ki jim pripada ta oznaka, je *drevo*. Za izdelavo oblikoskladenjskih označevalnikov obstaja več orodij in metod; za slovenščino sta pomembni predvsem naslednji dve:

- označevalnik podjetja Amebis (Amebis, 2011), ki temelji na ročno definiranih pravilih in leksikonu (učenje ni potrebno);
- označevalnik TnT (Brants, 2000), ki temelji na statistični analizi zaporedij besed in oznak v besedilu (potrebno je učenje na označenem korpusu). Na osnovi te tehnologije je bil naučen označevalnik, ki je del orodja TOTALE (Erjavec et al., 2005), nova različica pa je bila izdelana v okviru projekta JOS (Erjavec et al., 2010).

Označevalnik TnT je pravilneje označeval besede na podlagi empiričnega testiranja (Erjavec, 2010), in sicer s točnostjo (accuracy) $88.7\ \%$, medtem ko je Amebisov označevalnik na isti testni množici dosegel točnost $87.9\ \%$.

Zanimiv pa je bil tudi poskus izdelave metaoznačevalnika (Jan Rupnik and Miha Grčar and Tomaž Erjavec, 2010). Metaoznačevalnik[1] tvori iz dveh zaporedij oznak za isto besedilo eno ciljno zaporedje oznak. Z uporabo dveh neodvisnih označevalnikov (na primer Amebisovega in označevalnika TnT) dobimo začetni zaporedji oznak. V primerih, ko označevalnika nista enotna o določitvi oznake za določeno besedo, meta-označevalnik na podlagi naučenih pravil izbere eno od obeh pripisanih oznak, tako da poveča točnost končnega zaporedja oznak.

[1] Dostopen na strani: http://oznacevalnik.slovenscina.eu/Vsebine/Sl/ProgramskaOprema/Meta.aspx

2.1.1.3 Leme

Lema v oblikoslovju (lemma in morphology) predstavlja kanonično obliko nekega leksema. Leksem se v tem kontekstu nanaša na sklop vseh oblik neke besede, ki imajo enak pomen, lema pa se nanaša na besedo, ki je izbrana zato, da predstavlja leksem. Proces za določanje leme se imenuje lematizacija.

2.1.1.4 Krni

Krn v oblikoslovju (stem in morphology) predstavlja osnovno obliko besede. V Slovenski slovnici (Toporišič, 2000) je krn opisan kot besedotvorna podstava. Krnjenje (stemming) je jezikovno odvisen postopek, pri katerem poskušamo najti niz znakov, ki ga imenujemo krn in lahko predstavlja vse oblike neke besede ter istočasno to besedo loči od vseh ostalih. Pogosto, vendar ne nujno, krn ustreza korenu besede. Primer 2.1 kaže takšno razliko.

(2.1) *krožiti*
krož-*iti*
krn: krož
koren: krog

Krnjenje je še posebej pomembno pri avtomatskem indeksiranju besedil v jezikih z razvejanim oblikoslovjem, kakršna je tudi slovenščina.

2.1.1.5 Paradigme

Za potrebe pričujočega dela zadošča, da paradigme opišemo kot razred elementov z enakim obnašanjem. Paradigma je sestavljena iz pravil, ki vsaki oznaki MSD, dovoljeni v tej paradigmi, pripiše spremembo obrazil. Slika 2.1 kaže del paradigme s tipično predstavnico *žoga*. Končnica *a* se spreminja glede na MSD, primer orodnik: *a—>ami*.

```
paradigma: žog/a
a - samostalnik ženski množina imenovalnik
am - samostalnik ženski množina dajalnik
e - samostalnik ženski množina tožilnik
ami - samostalnik ženski množina orodnik
```

Slika 2.1: Del paradigme za samostalnike ženskega spola v slovenščini. Tipični predstavnik je lema *žoga*. Končnica -*a* se spreminja v skladu z različnimi MSD.

Paradigme, v našem primeru pregibne paradigme (inflectional morphology paradigm), so večdimenzionalne, potencialno rekurzivne matrike, ki so določene z oblikoslovnimi značilnostmi besednih oblik in obrazil (Clahsen et al., 2000). Teoretični status pregibnih paradigem ima nasprotujoče razlage: Lieber (1992), na primer, trdi, da so paradigme le skupine, podobne seznamom povezanih stavkov. Tudi Halle in Marantz (1993) predstavljata pregibne paradigme brez teoretičnega statusa. V večini drugih razlag, na primer (Wurzel, 1987) ali (Spencer, 1991), pa pregibna paradigma pomeni sklop pregibnih besednih oblik za vsak leksem, ki sodi v neko skladenjsko kategorijo.

Neki lemi lahko pripišemo paradigmo, ki vsebuje nabor vseh MSD, ki so dovoljeni za to lemo in kjer veljajo pravila pretvorbe obrazil za vse besedne oblike te leme.

Oglejmo si še en primer: porazdelitev samostalnikov na sklanjatvene vzorce kot jih predstavlja Toporišič (2000) omogoča izdelavo zbirke pravil, ki omogočajo natančno rokovanje z veliko množico besed (besede razdelimo na krne in s končnicami izbiramo ostale jezikovne kategorije, kot so spol, število, sklon). Povezava med posamezno besedno obliko in njeno paradigmo poteka prek osnovne besedne oblike – leme. Tak način opisa besed prinaša poseben problem samodejnega označevanja parov lema-paradigma, ki je opisan v razdelku 4.3.1.1.

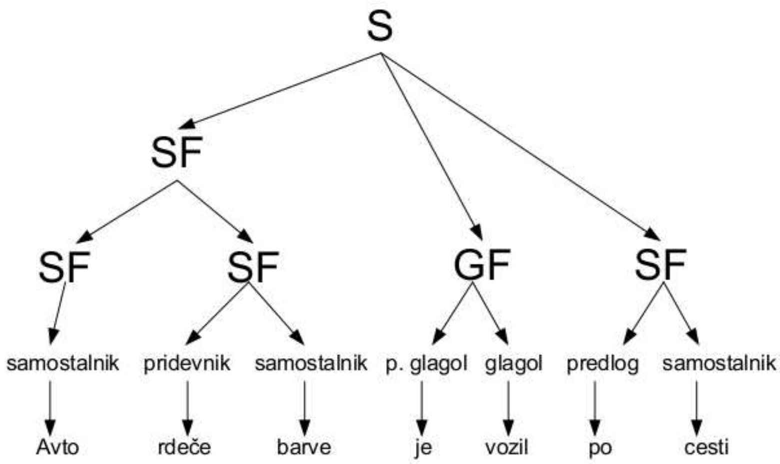
2.1.2 Drevesa izpeljav

Drevo izpeljav (parse tree ali concrete syntax tree) je urejeno drevo, ki predstavlja sintaktično strukturo niza glede na neko (formalno) slovnico. Sestavljeno je iz korena (vrhnjega vozlišča), vej drevesa, notranjih vozlišč in listov (končnih vozlišč). Pri drevesu izpeljav so notranja vozlišča označena z neterminalnimi simboli (non-terminals) slovnice, listi drevesa, končna vozlišča pa s terminalnimi simboli slovnice. Drevesa izpeljav lahko izdelamo za povedi naravnih jezikov glede na njihove slovnice. Slika 2.2 kaže primer jezikoslovnega drevesa izpeljav za preprost slovenski stavek *„Avto rdeče barve je vozil po cesti"*.

2.1.3 Plitko skladenjsko razčlenjevanje in plitki prenos besedil

Plitko skladenjsko razčlenjevanje besedil (glsshallow parsing, glschunking, glslight parsing) je postopek razčlenjevanja, ki določa osnovne gradnike povedi, samostalniške skupine, glagole, glagolske skupine itd., vendar pa ne navaja notranje zgradbe gradnikov niti njihove vloge v povedi.

Naravno nadaljevanje plitkega razčlenjevanja je pri postavitvi prevajalnega sistema plitki prenos, saj omogoča prenos gradnikov plitkega razčlenjevanja izvorne



Slika 2.2: Drevo izpeljav za stavek *Avto rdeče barve je vozil po cesti.* S – stavek, SF – samostalniška fraza, GF – glagolska fraza.

povedi v gradnike ciljnega jezika. Plitki prenos je sestavljen iz prenosa posameznih besed, po navadi v lematizirani obliki (glej razdelek 2.1.1.3), in pravil plitkega prenosa, ki služijo za opis sprememb med izvornim in ciljnim jezikom. Ta pravila so omejena na lokalne spremembe, kot sta na primer ujemanje soležnih besed v leksikalnih kategorijah in sprememba lokalnega vrstnega reda besed. Primeri pravil so prikazani v prilogi A.

2.1.4 Morfemi

Morfem je najmanjši del besede s samostojnim pomenom. Morfeme na izrazni ravni sestavljajo fonemi, najmanjše enote govorjenega jezika, s katero govorci določenega jezika razlikujejo pomen besed. V pisni obliki pa so morfemi sestavljeni iz grafemov, najmanjših enot pisnega jezika. Slika 2.3 kaže primer razdelitve besede *obkrožim* na morfeme, razložen je tudi pomen morfemov.

2.1.5 Besedni razredi

Besedni razredi definirajo kategorije besed. Kategorije so definirane s pogledom na obnašanje besed, ki pripadajo določenemu razredu. Primera takšnih razdelitev besednih razredov so oznake PoS in MSD.

```
obkrožim
ob - zraven, okoli
krož - krožiti
i - dovršnost
m - določenost glede na delo
```

Slika 2.3: Razdelitev besede *obkrožim* na morfeme in razlaga pomena posameznih morfemov.

Pri statistični obdelavi naravnih jezikov (natural language processing) se pogosto srečujemo s problemom redkih podatkov (sparse data problem). Eden od običajnih načinov reševanja tega problema je razvrščanje besed v ekvivalenčne razrede, besedne razrede (word classes).

2.1.6 Oblikoskladenjska analiza besedil

Oblikoskladenjska analiza je proces razčlenjevanja besed na njihove morfeme (pomenske enote). Oblikoskladenjska analiza je bistvena komponenta aplikacij jezikovnih tehnologij, uporabna pri odkrivanju pravopisnih napak (spelling error correction), strojnem prevajanju in drugem. Izvajanje polne oblikoskladenjske analize besedila običajno zahteva delitev besed na morfeme in analizo interakcije teh morfemov. Obe aktivnosti določata skladenjske razrede besednih oblik kot celote. Kompleksnost oblikoskladenjske analize se med naravnimi jeziki močno razlikuje, velja pa za relativno težek problem že v relativno preprostih primerih, kot je angleščina.

2.1.7 Pravila prenosa na osnovi regularnih izrazov

Plitki strukturni prenos (shallow structural transfer) omogoča premostitev slovničnih razlik obravnavanega jezikovnega para. Temelji na tehnologiji končnih avtomatov za odkrivanje vzorcev leksikalnih enot (oblikoskladenjsko označenih delov besedila ali fraz) konstantne dolžine, ki zahtevajo posebno obdelavo glede na slovnične razlike med jezikoma (na primer: spremembe v spolu, sklonu ali številu za zagotovitev ujemanja v ciljnem jeziku).

2.1.8 Statistični modeli jezika

Cilj statističnega jezikovnega modeliranja (statistical language modelling) je izdelava statističnega modela, ki omogoča oceno distribucije naravnega jezika. Statistični model jezika (statistical language model – SLM) predstavlja distribucijo verjetnosti besednih nizov, ki odseva, kako pogosto se določen niz besed pojavlja v jeziku.

Jezikovno modeliranje se uporablja v mnogih vrstah jezikovnih tehnologij, kot so prepoznavanje govora, strojno prevajanje, označevanje MSD, razčlenjevanje in priklic informacij (information retrieval).

Najpogostejše tehnike jezikovnih modelov so:

- Modeli, temelječi na n-gramih, n-gramski modeli (n-gram language model), so najpogosteje uporabljani statistični jezikovni modeli. Verjetnost niza besed $P(S)$ predstavimo s formulo v enačbi 2.1, v kateri z w_i označimo i-to besedo niza S. Torej je verjetnost niza besed S enaka produktu pogojnih verjetnosti posameznih besed, ki sestavljajo S, pri pogoju, da pred i-to besedo v nizu nastopajo vse predhodne besede S.

$$\begin{aligned} P(S) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...p(w_l|w_1...w_{l-1}) = \\ = \prod_i^l P(w_i|w_1...w_{i-1}) \end{aligned} \tag{2.1}$$

  Osnovne parametre modela izračunamo s tehniko ocenjevanja največjega verjetja (maximum likelyhood estimation – MLE), ki je predstavljena z enačbo 2.2, v kateri funkcija C pomeni štetje pojavitev (Count), torej z enostavnim štetjem pojavitev posameznih besed v učni množici, učnem korpusu.

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \tag{2.2}$$

  Modeli, temelječi na n-gramih, so najširše raziskani, predlaganih je tudi več razširitev, kot so Class-based N-gram model (Brown et al., 1992), Grammatical Trigrams (Lafferty et al., 1992). Najpogosteje je uporabljen model z dolžino n-grama 3, trigramski model.

- Model največje entropije (maximum entropy language model) je splošen statistični model, ki lahko vsebuje funkcije iz različnih virov. Pogojni ME-model ima obliko, kot jo kaže enačba 2.3:

$$p(w|h) = \frac{1}{Z(h)} e^{\sum_i \lambda_i \phi_i(h_i, w)}, \tag{2.3}$$

v kateri so λ parametri, w je beseda, h je zgodovina, $Z(h)$ je normalizacijski faktor in $\phi(h, w)$ so poljubne funkcije besede in zgodovine. Model ME, ki vključuje trigrame, oddaljenost N-gramov in model na osnovi parov sprožil (trigger pairs), je bil primerjan z osnovnim trigramskim modelom. Ugotovljeno je bilo $30\ \%$ zmanjšanje nivoja začudenja (perplexity) v primerjavi z izhodiščnim trigramskim modelom (Rosenfeld, 1996).

- Jezikovni modeli s strukturnimi podatki (Structured Language Models) upoštevajo tudi strukturo povedi. Jezikovni modeli na osnovi n-gramov ne morejo opisati oddaljenih odvisnosti (long-distance dependencies) v podatkih. Predlagani so bili različni alternativni pristopi modeliranja jezika, ki vključujejo strukturne podatke za modeliranje jezika. Primeri takšnih modelov so:
  - Jezikovni modeli na osnovi sprožil (trigger models) (Raymond et al., 1993) kot osnovno jezikovno informacijo uporabljajo pare sprožil. Če je neko besedno zaporedje *A* značilno korelirano z besednim zaporedjem *B*, potem *AB* imenujemo sprožilni par. Če se zaporedje *A* pojavi v besedilu, sproži spremembo ocene verjetnosti za *B*.
  - Jezikovni modeli s preskakovanjem n-gramov (skipped n-gram language models) (Rosenfeld, 1994). Modeli temeljijo na podobnem postopku izdelave n-gramov kot pri osnovnih modelih temelječih na n-gramih, vendar poleg zaporedja sosednjih besed uporabljajo tudi možnost "preskočitve"besede, torej določene besede ne upoštevamo pri gradnji n-grama. Za opisovanje oddaljenih odvisnosti bi lahko uporabili večje n-grame, vendar se z njihovo velikostjo manjša verjetnost, da v učnih podatkih obstaja podoben kontekst. Če pa uporabimo le večino besed, dobimo modele s preskakovanjem n-gramov.
  - Jezikovni model s skladenjskim razčlenjevanjem (LM by Syntactic Parsing) za modeliranje jezika uporablja verjetnost razčlenjevalnika, kot je na primer (Charniak, 1997).

V nadaljevanju je v vseh predstavitvah za statistični jezikovni model uporabljen trigramski jezikovni model (trigram language model). Za uporabo v metodah, ki so predstavljene v tem delu, model na osnovi trigramov zadošča in uporaba zapletenejših modelov ni smiselna. Izdelava modela je natančneje predstavljena v razdelku 4.3.3.

2.1.9 Končno urejanje

Končno urejanje (post-editing) je postopek za izboljšanje strojno ustvarjenih prevodov s čim manj ročnega dela. Vključuje popravljanje strojno izdelanih prevodov za zagotovitev ravni kakovosti, ki je bila dogovorjena med stranko in prevajalcem. Določene ponovljive operacije lahko avtomatiziramo oziroma vključimo v same strojne prevajalnike. Primer 2.2 kaže združevanje določnega člena in predloga v italijanščini v predložno zvezo (preposizione articolata).

(2.2)	*stretta*	*di*	*la*	*mano*
	stretta	di-*določni člen*	la-*predlog*	mano
	"(stretta) di la mano"			
	"(stisk) roke"			

stretta	*della*	*mano*
stretta	della-*predložna zveza*	mano
"stretta della mano"		
"(stisk) roke"		

2.2 Slovanski jeziki

Slovanski jeziki so velika jezikovna družina v Srednji in Vzhodni Evropi ter na Balkanu in v delu Azije. Po številu govorcev je največji slovanski jezik ruščina, sledi poljščina.

Večina slovanskih jezikov, razen bolgarščine in makedonščine, ima razvito bogato pregibanje samostalnikov.

Beloruščina je jezik, ki ga govori približno 7 milijonov ljudi v Belorusiji. Spada v skupino vzhodnoslovanskih jezikov ter je soroden ruščini in ukrajinščini.

Bolgarščina je jezik, ki ga govori približno 7 milijonov ljudi v Bolgariji. Velja za poseben jezik med slovanskimi jeziki, saj je izgubil samostalniške sklanjatve. Soroden jezik bolgarščini je makedonščina.

Bosanščina, hrvaščina, črnogorščina in srbščina (Bosnian, Croatian, Serbian – BCS) so jeziki, ki jih govorijo na območju nekdanje Jugoslavije. Sodijo v južnoslovanski narečni kontinuum. V preteklosti je bil za te jezike uporabljen skupni izraz srbohrvaščina. So sorodni slovenščini na severozahodu ter bolgarščini in makedonščini na jugovzhodu. Skupaj imajo ti jeziki okoli 16 milijonov govorcev.

Češčina je zahodnoslovanski jezik, ki ga govori približno 10 milijonov ljudi v Češki republiki. Zaradi zgodovinskih okoliščin obstajata dve različici jezika, knjižna in pogovorna, med katerima obstajajo velike razlike. Poleg ruščine je češčina slovanski jezik z največ računalniškimi jezikovnimi viri in orodji.

Kašubščina ali pomorjanščina je jezik, ki ga govori približno 50.000 ljudi na severu Poljske. Vsi Kašubi so dvojezični (Poljska). Jezik je soroden slovinskemu jeziku (tudi v Severni Poljski), ki je izumrl na začetku 20. stoletja.

Makedonščina je južnoslovanski jezik, ki ga uporablja približno 1,5 milijona ljudi v Makedoniji ter makedonske manjšine v Albaniji, Bolgariji in Egejski Makedoniji (današnja Grčija). Sorodna je bolgarščini.

Poljščina je zahodnoslovanski jezik, ki ga govori približno 38 milijonov ljudi na Poljskem in nacionalne manjšine v Belorusiji, Republiki Češki, Litvi in Ukrajini.

Ruščina je vzhodnoslovanski jezik, ki ga uporablja približno 150 milijonov ljudi v Rusiji in nekdanjih sovjetskih republikah. Je slovanski jezik z največ govorci. Ena od zanimivih lastnosti ruščine je, da nima preteklika, ki bi uporabljal pomožni glagol.

Slovaščina je zahodnoslovanski jezik s približno 4,5 milijona govorci. Je del češko-slovaškega narečnega kontinuuma (Townsend in Janda, 2003) in je soroden češčini, razlike so predvsem na fonetični ravni.

Slovenščina je južnoslovanski jezik, ki ga govorijo Slovenci v Sloveniji ter narodne manjšine na avstrijskem Koroškem, v Italiji in na Madžarskem. Uporablja ga približno 1,8 milijona govorcev. Jezik je ohranil nekaj arhaičnih značilnosti, kot na primer dvojino.

Spodnjelužiška srbščina je jezik, ki ga govori približno 15 milijonov ljudi v nemški deželi Spodnja Lužica. Kot obrobna narečja ohranja veliko starih jezikovnih značilnosti, kot so dvojina in veliko preteklih časov (aorist in imperfekt). Po drugi strani pa je bil pod močnim vplivom jezika okolja, nemškega jezika.

Ukrajinščina je vzhodnoslovanski jezik, ki ga uporablja približno 37 milijonov ljudi v Ukrajini. Podobno kot poljščina je ohranila pasivni pretekli deležnik.

Zgornjelužiška srbščina je jezik, ki ga govori približno 35.000 ljudi v nemški deželi Zgornja Lužica. Za ta jezik veljajo podobne lastnosti kot za Spodnje lužiško srbščino, ki je opisana v tem razdelku.

Starocerkvena slovanščina je izumrli jezik, v katerem so bila napisana najstarejša, predvsem bogoslužna slovanska besedila. Temelji na srednjeveškem narečju makedonske metropole Solun in je bil jezik bogoslužja Velike Moravske. Jezik je dobro dokumentiran, obstajajo pisane slovnice in slovarji. Večina modernejših jezikov je z leti izgubljala posamezne značilnosti, ki jih je imela starocerkvena slovanščina.

Polabščina je izumrli zahodnoslovanski jezik, ki se je uporabljal v severovzhodni Nemčiji, natančneje med spodnjo ter srednjo Labo na zahodu ter spodnjo Odro na vzhodu. Izumrl je v 18. stoletju. Polabščina je sorodna kašubščini in lužiški srbščini.

2.3 Podobnosti slovanskih jezikov kot pomoč pri prevajanju

Pri izdelavi sistemov za strojno prevajanje sorodnih jezikov izkoriščamo podobnosti med jezikoma jezikovnega para (free rides). Izkušnje s področja strojnega prevajanja med sorodnimi jeziki (Homola, 2010) kažejo, da je smiselno razdeliti podobnosti jezikov na kategorije (nivoje) ujemanja. Ločimo tipološko, oblikoslovno, skladenjsko in leksikalno podobnost. V nadaljevanju sledi pregled posameznih kategorij z vidika strojnega prevajanja.

2.3.1 Tipološke podobnosti

Za namene strojnega prevajanja je najpomembnejša tipološka kategorija podobnosti. Če sta jezika prevajanega jezikovnega para iz različnih tipoloških skupin, je prevajanje oteženo. Funkcije, kot so besedni vrstni red, obstoj oziroma neobstoj členov, različni sistemi časov in podobne razlike, predstavljajo za strojno prevajanje najhujše prepreke.

Oglejmo si primer slovenščine in makedonščine kot jezikov, ki pripadata isti jezikovni skupini, a se tipološko razlikujeta. Podoben primer bi lahko predstavili tudi za češčino ali srbščino ter makedonščino. Oba jezika poznata bogato pregibanje glagolov in načeloma prost besedni vrstni red, zato ni treba spreminjati vrstnega reda glagolov. Velika razlika med jezikoma pa je v dejstvu, da makedonščina ne pozna samostalniških sklanjatev.

Primera 2.3 in 2.4 pomenita približno isto, in sicer „Moj brat je bral knjigo”. Tabela 2.1 opisuje kratice oblikoskladenjskih značk, ki so uporabljene v primerih. Večina kratic je iz nabora specifikacije JOS (Erjavec et al., 2009), kratice, označene z $*$, so po specifikaciji MULTEXT-EAST (Erjavec, 2004).

Tabela 2.1: Razširjene kratice, ki so uporabljene v Primerih 2.3 in 2.4. Večina kratic je iz nabora specifikacije JOS, kratice, označene z $*$, so po specifikaciji MULTEXT-EAST.

Somei	samostalnik občno_ime moški ednina imenovalnik
Soset	samostalnik občno_ime srednji ednina tožilnik
Sozet	samostalnik občno_ime ženski ednina tožilnik
Sozei	samostalnik občno_ime ženski ednina imenovalnik
Gp-ste-n	glagol pomožni sedanjik tretja ednina nikalnost
Ggnd-em	glagol glavni nedovršni deležnik ednina moški
Gp-ppe-n	glagol pomožni prihodnjik prva ednina nikalnost(ne)
Ggdd-em	glagol glavni dovršni deležnik ednina moški
Ggdn	glagol glavni dovršni nedoločnik
Gp-spe-n	glagol pomožni sedanjik prva ednina nikalnost(ne)
Pp2-sa–n*	pronoun personal second singular accusative clitic=no
Zop-ed	zaimek osebni prva ednina dajalnik
Zspmeie	zaimek svojilni prva moški ednina imenovalnik ednina
Rsn	prislov splošni nedoločeno
Vmii3s*	verb main imperfect tense indicative singular third person
Npfsny*	noun proper feminine singular nominative Definiteness=yes

(2.3) *Moj brat je bral knjigo.*
Zspmeie Somei Gp-ste-n Ggnd-em Sozet.
“Moj brat je bral knjigo.”

Брат ми читаше книга.
Somei Zop-ed Vmii3s Sozei
“Brat mi čitaše kniga.”

(2.4) *Knjigo je bral moj brat.*
Sozet Gp-ste-n Ggnd-em Zspmeie Somei .
”Knjigo je bral moj brat.”

Книгата	ја	читаше	брат	ми.
Npfsny	*Pp2-sa–n*	*Vmii3s*	*Somei*	*Zop-ed*

"Knigata ja čitaše brat mi."

Zaradi skoraj prostega besednega vrstnega reda je pomen v obeh primerih enak. V angleščini bi tako tvorili pasivno obliko: *My brother read a book* in *The book has been read by my brother*. V makedonščini je besedni vrstni red enak kot v slovenščini.

2.3.2 Skladenjske podobnosti

Skladenjska podobnost je pomembna predvsem v povezavi z glagoli. Razlike v glagolski vezavi (verb valency) negativno vplivajo na kakovost prevoda, saj v fazi prenosa zahtevajo uporabo vezljivostnih slovarjev (valency dictionaries) izvornega in ciljnega jezika. Izdelava takšnih slovarjev je zapletena in predvsem draga. Razlike v skladenjskih strukturah manjših sestavin, kot so samostalniške in predložne besedne zveze, na kakovost prevodov nimajo takšnega vpliva. Analiza takšnih struktur je možna s pomočjo plitkega skladenjskega razčlenjevanja, sprememba skladenjske strukture ciljne povedi je lokalnega značaja.

Za sorodne jezike po navadi velja, da se besedni vrstni red v prevodu ne spreminja. Obstajajo tudi izjeme, kot kaže primer 2.5: pri prevodih povedi v prihodnjiku med slovenščino in srbščino ter slovenščino in hrvaščino se besedni vrstni red spremeni.

(2.5)

Jaz se bom oblekel. (SLO)

Ja ću da se obučem. (SR)

Ja ću se obući. (CR)

2.3.3 Oblikoslovne podobnosti

Oblikoslovna (morfološka) podobnost pomeni podobno strukturo oblikoslovne hierarhije in paradigem, kot na primer podobnosti v sistemu sklonov, podobnosti pri spreganju glagolov itd. Slovanski jeziki, z izjemo makedonščine in bolgarščine, imajo podobne sklanjatvene in spregatvene vzorce. Razlike v oblikoslovju lahko razmeroma enostavno odpravimo z izkoriščanjem polnih oblikoslovnih modulov za oba jezika jezikovnega para. Podobni oblikoslovni sistemi lajšajo fazo prenosa. Na primer, večina slovanskih jezikov, razen bolgarščine in makedonščine, pozna 6 ali 7 sklonov.

Nekaj problemov povzročajo sintetične forme, ki zahtevajo analitične konstrukte v drugih jezikih. Tak je primer prevajanja prihodnjika med slovenščino in srbščino. Primer 2.6 kaže prevod slovenske povedi v prihodnjiku v srbsko poved. Pomožni glagol *biti* v prihodnjiku pri prevodu spremeni lemo v *hteti* ter čas v sedanjik, glavni glagol, v tem primeru *kupiti*, pri prevodu iz preteklika preide v nedoločnik.

(2.6) *Jutri* *bom* *kupil* *darilo.*
jutri-*Rsn* biti-*Gp-ppe-n* kupiti-*Ggdd-em* darilo-*Soset*
"Jutri bom kupil darilo." (SLO)

Sutra *ću* *kupiti* *poklon.*
sutra-*Rsn* hteti-*Gp-spe-n* kupiti-*Ggdn* poklon-*Soset*
"Sutra ću kupiti poklon." (SR)

Razlike, kot je prikazana na primeru 2.6, rešujemo s pomočjo pravil za plitki prenos. Pri prevodu se zamenja lema pomožnega glagola izvornega jezika *biti* v lemo ciljnega jezika *hteti* ter oblika glagola v ciljnem jeziku v nedoločnik (v prikazanem primeru glagola *kupiti*).

Obširneje so primeri pravil prikazani v prilogi A.

2.3.4 Leksikalne podobnosti

Leksikalna podobnost ne pomeni, da mora imeti besedišče enak izvor (enako etimologijo), da morajo besede izvirati iz istega korena. Kar je pomembno za strojno prevajanje, je semantično ujemanje besed, po možnosti ena-na-ena, torej za vsako izvorno lemo obstaja le ena ciljna lema in obratno.

Leksikalna podobnost je z vidika strojnega prevajanja najmanj pomembna, z drugimi besedami, leksikalne razlike pri prevajanju enostavno premoščamo z uporabo glosarjev in splošnih slovarjev.

Kljub temu pa bo morda treba dvojezične slovarje razširiti z oblikoskladenjskimi podatki. Oglejmo si takšen primer na jezikovnem paru slovenščina-srbščina; obstaja namreč nekaj samostalnikov, ki so različnih spolov v obeh jezikih. Primer 2.7 kaže spremembo spola iz srednjega v moški pri prevodu besede *okno* v srbsko besedo *prozor*. V obeh jezikih se pridevnik ujema s samostalnikom v spolu.

(2.7) *Odprto* *okno.*
odprto-*pridevnik, srednji spol* okno-*samostalnik, srednji spol*
"Odprto okno." (SLO)

Otvoren	*prozor.*
otvoren-*pridevnik, moški spol*	prozor-*samostalnik, moški spol*
"Otvoren prozor." (SR)	

To razliko je mogoče popraviti med leksikalnim prenosom. V tej fazi ciljna lema zamenja mesto izvorne in obdrži ostale oblikoskladenjske oznake. Oznako za spol zamenjamo z ustrezno oznako ciljne leme. Takšen popravek pa povzroči neujemanje spremenjenega samostalnika z okoliškimi besedami. V večini slovanskih jezikov se sosednja samostalnik in pridevnik ujemata v spolu, sklonu in številu, v nekaterih primerih tudi v drugih oblikoskladenjskih kategorijah. Problem rešujemo s pravili za lokalno ujemanje, s katerimi se popravljajo porušena lokalna ujemanja v oblikoskladenjskih kategorijah. Pravilo lokalnega ujemanja samostalnika in pridevnika je predstavljeno na sliki 5.11 v razdelku 5.5.

2.3.5 Podobnice

Podobnice (cognates) so besede, ki imajo skupen etimološki izvor. Pri prevajanju med dvema jezikoma so predvsem dobrodošle podobnice, ki se s časom niso veliko spremenile ne v pomenu niti v obliki. Takšne besede lahko med jezikoma jezikovnega para prevajalnega sistema prevajamo z malenkostnimi spremembami (po navadi končnice ali posamezne črke). V tabeli 2.2 je nekaj primerov parov podobnic med različnimi jezikovnimi pari.

Pri uporabi podobnic moramo paziti na *lažne podobnice*, besede, ki so si podobne, a imajo drugačen etimološki izvor, in na *lažne prijatelje*, podobne besede z različnim pomenom. Uporaba prvih, oziroma zanašanje na podobnosti pri prvih primerih, za samo prevajanje ni problematična, saj nas sicer napačno povezovanje besed po izvoru kljub temu pripelje do pravilnih prevodov. Lažni prijatelji pa povzročajo napačne prevode, ki so natančneje razloženi v razdelku 2.3.6.

2.3.6 Lažni prijatelji

Lažni prijatelji (false friends) so po obliki podobne besede v različnih jezikih, ki imajo različne pomene. Problem lažnih prijateljev je še zlasti opazen pri sorodnih in sosednih jezikih, predvsem pri površnih poznavalcih jezikovnih parov. Lažni prijatelji lahko povzročajo težave pri učenju tujih jezikov, predvsem jezikov sorodnih maternemu. Lažnih prijateljev je med slovanskimi jeziki veliko, primeri so predstavljeni v tabeli 2.3.

Tabela 2.2: Primeri podobnic: na začetku vsake vrstice je slovenski pomen podobnic, ki sledijo. Spisek izbranih jezikov uvaja primere podobnic v naslednji vrstici.

SLO pomen	slovenščina	hrvaščina	srbščina
mleko zvezda noč	mleko zvezda noč	mlijeko zvijezda noč	mleko zvezda noč
SLO pomen	italijanščina	španščina	portugalščina
mleko zvezda	latte stella	leche estrella	leite estrela
SLO pomen	angleščina	nemščina	švedščina
mleko zvezda	milk star	milch sterne	mjölk stjärna

Pri izdelavi sistemov za strojno prevajanje moramo biti pazljivi zlasti v fazi gradnje dvojezičnih prevajalnih leksikonov. Z metodami, ki slonijo le na oblikovni podobnosti besed, lahko napačno uporabimo lažne prijatelje.

Tabela 2.3: Primeri lažnih prijateljev, to je podobnih besed v različnih jezikih z različnimi pomeni.

primer	pomen	jezik	primer	pomen	jezik
tanjši	tanjši	SLO	tanji	cenejši	poljščina
najlepši	najlepši	SLO	nejlepši	najboljši	češčina
drago	drag	SLO	drago	ljubo	hrvaščina/srbščina
prost	prost	SLO	prost	preprost/vulgaren	hrvaščina/srbščina

2.4 Uporabljena učna gradiva

Metode za hitro izdelavo sistemov za strojno prevajanje naravnih jezikov, ki so predstavljene v nadaljevanju dela, temeljijo na učnih gradivih, iz katerih se izlušči znanje o jeziku in predvsem znanje o razlikah med jezikoma, ki ga sistem za prevajanje potrebuje pri samih prevodih. Uporabili smo večjezični oblikoskladenjsko označeni korpus relativno majhnega obsega ter večje, enojezične in neoznačene korpuse.

2.4.1 Korpus Multext-east

Pri večini metod, predstavljenih v razdelku 4.3, je bil kot učna množica uporabljen večjezični poravnani korpus MULTEXT-EAST[2] (Erjavec, 2010; Dimitrova et al., 1998), ki je večjezična zbirka jezikovnih gradiv. Zapisan je v standardizirani obliki in podpira velik del srednje- in vzhodnoevropskih jezikov. Uporablja oblikoskladenjske označbe po vzoru EAGLES (Leech in Wilson, 1996). Korpusni del gradiv je zapisan v standardizirani obliki v formatu XML (Bray et al., 2008) po smernicah TEI-P5 (TEI-Consortium, 2007). Gradiva sestavljajo oblikoskladenjske specifikacije, oblikoskladenjski leksikoni ter označeni, vzporedni, primerjalni in govorni korpusi. Trenutna različica gradiv obsega 16 jezikov in je prosto dostopna za raziskovalne namene.

Tabela 2.4 kaže število različnih besednih oblik in število lem za izbrane jezike, ki smo jih uporabili v raziskavah, predstavljenih v tem delu.

Tabela 2.4: Število lem in besednih oblik za slovenščino, češčino, srbščino, angleščino in estonščino

jezik	število besednih oblik	število lem
slovenščina	20.923	7.895
srbščina	2. 505	8.392
češčina	22.273	9.060
angleščina	11.078	7.020
estonščina	18.853	8.679

Primer povedi iz korpusa je prikazan na sliki 2.4; vsaka poved je shranjena v znački *s*, atribut *id* služi za povezavo z drugimi jeziki. Vsaka beseda je shranjena v znački *w*, atribut *lemma* predstavlja lemo besede, atribut *ana* pa oblikoskladenjsko oznako besede. Značka *c* označuje ločila.

Vzporedni del korpusa, ki je uporabljen v metodah, predstavljenih v nadaljevanju, sestavlja roman Georga Orwella „1984” (Orwell, 1949), preveden v vseh 16 jezikov. Celoten roman je oblikoskladenjsko označen in vzporedno poravnan na nivoju povedi s pivotnim jezikom – angleščino. Vsi prevodi so poravnani z angleškim izvirnikom. V nadaljevanju dela bo ta korpus imenovan „1984”.

2.4.2 Jezikoslovno označevanje slovenskega jezika

Projekt JOS, Jezikoslovno Označevanje Slovenskega jezika (Erjavec et al., 2010), prinaša označene korpuse slovenskega jezika in pridružene vire, namenjene spod-

[2]Korpus je dostopen na naslovu: http://nl.ijs.si/ME/V4/.

```
<s id="Osl.2.3.5.11">
  <w lemma="priti" ana="Vmps-dma">Prišla</w>
  <w lemma="biti" ana="Vcip3d--n">sta</w>
  <w lemma="do" ana="Spsg">do</w>
  <w lemma="podrt" ana="Afpnsg">podrtega</w>
  <w lemma="drevo" ana="Ncnsg">drevesa</w>
  <c>,</c>
  <w lemma="o" ana="Spsl">o</w>
  <w lemma="kateri" ana="Pr-nsl----a">katerem</w>
  <w lemma="on" ana="Pp3msd--y-n">mu</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="praviti" ana="Vmps-sfa">pravila</w>
  <c>.</c>
</s>
```

Slika 2.4: Označena poved v korpusu „1984" (Erjavec, 2010).

bujanju razvoja jezikovnih tehnologij za slovenski jezik. Rezultati vsebujejo oblikoskladenjske specifikacije JOS (definicija nabora oblikoskladenjskih oznak) (Erjavec et al., 2009), dva označena korpusa in dva spletna servisa. Ponovno naučen, na novih oznakah, je tudi označevalec oznak MSD; poleg prehoda na nove oznake se je še izboljšala tudi kakovost označevanja. Razviti viri so v celoti dostopni in licencirani z licencami Creative Commons (Coates, 2007). Metode, predstavljene v nadaljevanju, uporabljajo starejše oznake MULTEXT-EAST (Erjavec, 2004), saj je njihov razvoj potekal še pred izidom rezultatov in smernic projekta JOS. Dodatno težavo povzročajo razlike v označevanju oznak MSD, ki sicer omogočajo boljše označevanje slovenskega jezika. Razlike v označevanju ostalih jezikov lahko izdelavo prevajalnega sistema otežijo.

2.4.3 Enojezični korpusi člankov iz Wikipedie

Enojezični korpus je bil uporabljen kot učna množica za izdelavo trigramskega jezikovnega modela. V okviru poskusov so bili izdelani korpusi za vse ciljne jezike testnih prevajalnih sistemov, ki so predstavljeni v nadaljevanju dela. Korpusi so sestavljeni iz naključno izbranih člankov iz Wikipedije za vse uporabljene jezike[3]. Velikost korpusov je bila približno 15 milijonov besed za češčino in angleščino ter

[3]http://cs.wikipedia.org, http://en.wikipedia.org, http://et.wikipedia.org, http://sr.wikipedia.org

približno 7 milijonov za estonščino in srbščino.

2.4.4 Korpus SVEZ-IJS

Korpus SVEZ-IJS ACQUIS (Erjavec, 2006) vsebuje celotno evropsko zakonodajo v obliki poravnanih povedi v slovenskem in angleškem jeziku, ki tvorijo angleško-slovenski pomnilnik prevodov SVEZ ACQUIS. Ta pomnilnik prevodov je izdelala prevajalska skupina SVEZ (Služba vlade RS za evropske zadeve) med procesom prevajanja zakonodaje EU (ACQUIS) v slovenski jezik. Korpus je oblikoskladenjsko označen po smernicah MULTEXT-EAST (Erjavec, 2004), lematiziran in normaliziran. Je največji večjezični korpus s slovenskim jezikom, vsebuje okoli 270.000 povedi v vsakem jeziku, skupaj torej približno 10 milijonov besed. Povedi so slabo oblikovane, z veliko naštevanja in posebnostmi pravnega jezika. Ta korpus je bil uporabljen za učenje besednih poravnav sistema za statistično strojno prevajanje na osnovi dreves izpeljav, ki je natančneje predstavljen v 6. poglavju.

2.4.5 Korpus JRC-Acquis

Korpus JRC-Acquis (Steinberger et al., 2006) vsebuje celotno evropsko zakonodajo od leta 1950 do najnovejših gradiv v obliki poravnanih povedi. Korpus obsega 22 uradnih jezikov evropske skupnosti, trenutno manjka le irski jezik. Sestavljen je iz več kot 4 milijonov dokumentov in več kot 630 milijonov besed . Ta korpus je bil uporabljen kot testna množica pri samodejni metriki vrednotenja METEOR, vrednotenje ja natančneje opisano v razdelku 7.2.2.1.

2.4.6 Korpus člankov dvojezičnega časopisa

Pri razvoju in empiričnem vrednotenju metode za vrednotenje pravil strukturnega prenosa, predstavljene v razdelku 5.4.3, je bil uporabljen korpus, zgrajen iz člankov dnevnika El Periódico de Catalunya. Velikost korpusa je približno 50.000 povedi oziroma 2 milijona besed.

2.4.7 Jezikoslovno označevanje z orodjem TOTALE

Orodje za jezikoslovno označevanje besedil TOTALE (Erjavec et al., 2005) je bilo uporabljeno pri označevanju korpusov, kot so Acquis Communautaire (Erjavec et al., 2005), JOS (Erjavec et al., 2010) in SVEZ-IJS ACQUIS (Erjavec, 2006). Orodje združuje skupek orodij za jezikovno označevanje, in sicer tokenizacijo, oblikoskladenjsko označevanje ter lematizacijo.

Samo označevanje poteka v treh delih:

- *tokenizacija*: delitev besedil na besede in ločila,
- *oblikoskladenjsko označevanje*: dodajanje oblikoskladenjskih oznak besedam, TOTALE za to nalogo uporablja orodje TnT (Brants, 2000),
- *lematizacija*: pripisovanje osnovne slovarske oblike posameznim besedam, TOTALE za to nalogo uporablja orodje CLOG (Erjavec in Džeroski, 2004).

TnT in CLOG sta programa, ki se jezikovnih modelov naučita na vnaprej pripravljenih podatkih, korpusih. Orodje bi lahko uporabili tudi za postavitev celotnega prevajalnega sistema, saj lahko uspešno nadomesti oblikoskladenjsko analizo vhodnega besedila in z manjšimi spremembami tudi oblikoskladenjsko sintezo. Jezikovni modeli, ki so izdelani z opisanima orodjema, so težko berljivi in napake težko odpravljamo, torej sistem ni primeren za dodatno izboljšanje, pri čemer lahko uporabimo jezikovno znanje.

2.5 Prevajalni sistem Guat

Prevajalni sistem Guat (Vičič, 2009) (ime je dobil po majhnih ribah Gobiidae, ki živijo tudi v slovenskem morju) je bil zgrajen med razvojem metod, prikazanih v 4. in 5. poglavju. Sistem podpira jezikovni par slovenščina-srbščina. Metode so bile preverjene skozi več iteracij (sistematične napake so bile popravljene in popravki so vključeni v osnovno ogrodje). Jezikovni par slovenščina-srbščina je bil uporabljen za preverjanje kakovosti predstavljenih metod na popolnoma delujočem prevajalnem sistemu. Posebnosti jezikovnega para so: oba jezika sta zelo pregibna, oblikoslovno in derivacijsko bogata. Čeprav sta jezika sorodna, visoka stopnja pregibnosti zahteva oblikoskladenjsko analizo izvornega jezika in posledično oblikoskladenjsko sintezo v končni fazi v ciljnem jeziku.

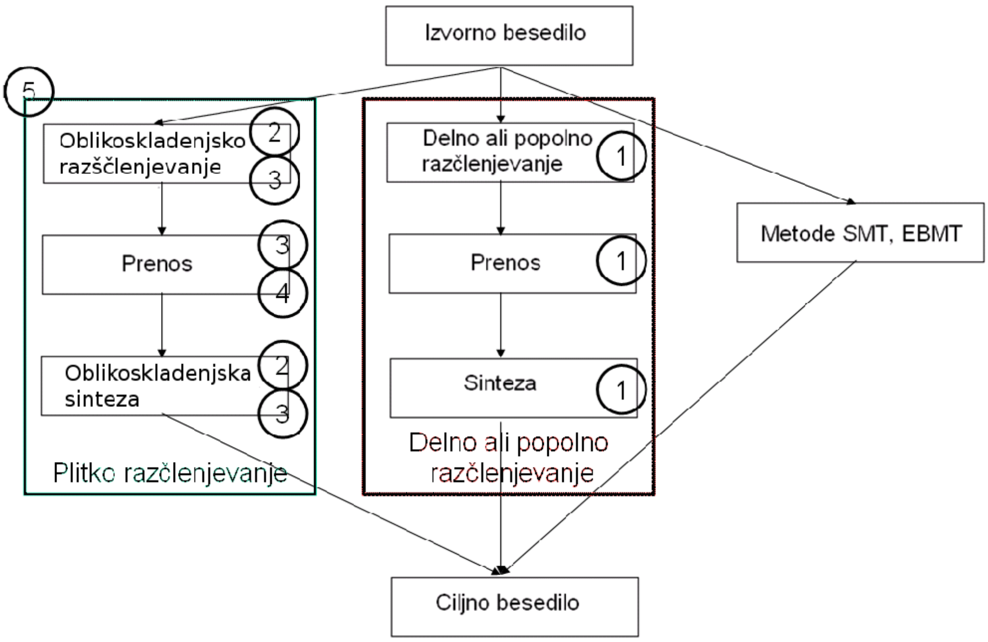
Prevajalni sistem je prosto dostopen na spletnem naslovu[4].

Eden od najprivlačnejših razlogov za uporabo sistema za strojno prevajanje na osnovi pravil (RBMT) je možnost, da strokovnjaki (jezikoslovci) izboljšajo samodejno izdelane jezikovne podatke. Strokovnjaki lahko enostavno dodajajo prevajalna previla in popravljajo napake v oblikoskladenjsko označenih slovarjih sistema ter v dvojezičnih prevajalnih slovarjih.

[4] Sistem je dostopen na spletnem naslovu: `http://jt.upr.si/guat/`.

2.6 Umestitev pričakovanih prispevkov k znanosti

Slika 2.5 kaže eno od možnih razdelitev strojnega prevajanja. V ospredju so metode, ki slonijo na pravilih (Rule Based Machine Translation – RBMT). Pričujoče delo se večinoma posveča prevajalnim sistemom te skupine. Predstavljeni so še alternativni sistemi, ki jih lahko v grobem združimo v dve skupini: statistično strojno prevajanje (Statistical Machine Translation – SMT) in strojno prevajanje na osnovi primerov (Example Based Machine Translation – EBMT). Sisteme, sloneče na pravilih, nadalje razdelimo na sisteme popolnega razčlenjevanja, kot sta Promt (Promt, 2010) in Systran (Systran, 2010), ter sisteme plitkega razčlenjevanja; primera takšnih sistemov sta Apertium (Corbi-Bellot et al., 2005) ter Česilko (Homola in Kuboň, 2008a).



Slika 2.5: Ena od možnih razdelitev strojnega prevajanja z umestitvijo pričakovanih prispevkov k znanosti. Prispevki so predstavljeni z zaporednimi številkami.

Prispevki k znanosti so oštevilčeni. Na sliki 2.5 so številke metod zapisane ob opisih posameznih delov prevajalnih sistemov. Prispevki so predstavljeni v nadaljevanju.

1. **Metoda za statistično strojno prevajanje z drevesi izpeljav za manj uporabljene jezike.** Metoda omogoča izdelavo sistema za statistično strojno pre-

vajanje z drevesi izpeljave za manj uporabljene jezike (less-used languages), ki nimajo izdelane drevesnice (treebank), tj. standardizirane zbirke skladenjsko označenih dreves, ki je običajna učna množica sistemov tega področja. Metoda je obširneje predstavljena v 6. poglavju.

2. **Metoda za samodejno označevanje paradigem.** Metoda je primerna za dodajanje novih besed v oblikoskladenjsko označen enojezični slovar. Predpogoj za metodo je dovolj dober in dovolj velik neoznačen enojezični korpus ter že izdelan oblikoskladenjsko označen slovar s paradigmami.

   Pravilna izbira paradigme omogoča enostavno izdelavo vseh besednih oblik nove leme. Besedi s pomočjo krnenja (Popovič in Willett, 1992) in dodajanja končnic vzorcev iz paradigem dodelimo množico z dvojniki (multiset) možnih pregibanj. Verjetnosti za posamezne paradigme izračunamo s preiskovanjem korpusa oziroma besed v njem. Paradigme z najvišjo oceno so dodatno preverjene s pragovno funkcijo. Metoda je obširneje predstavljena v 4. poglavju.

3. **Samodejno luščenje paradigem za visoko pregibne jezike in izdelava pripadajočih leksikonov.** Samodejno luščenje paradigem iz označenega korpusa predstavlja poseben primer razvrščanja neznanih primerov v množice in združevanj množic. Na osnovi vnaprej označenega enojezičnega korpusa je razvit algoritem za samodejno luščenje osnovnih paradigem. Paradigme so dodatno združene na osnovi podobnosti. Za preizkus metode je bil uporabljen korpus MULTEXT-EAST (Erjavec, 2010) in pripadajoči leksikoni. Metoda je obširneje prikazana v 4. poglavju.

4. **Ocenjevanje pravil za strukturni prenos.** Raziskava ocenjevanja pravil strukturnega prenosa je razdeljena na tri področja:

   - Raziskava možnih uporab ocenjevanja pravil. Porajajo se naslednje možnosti uporabe ocenjevanja pravil:

     – Ocenjevanje obstoječih pravil, ki so jih ročno pripravili strokovnjaki. To so pravila, ki se uporabljajo v sedanjih prevajalnih sistemih, temelječih na ogrodju Apertium. To ocenjevanje nam omogoča vzpostavitev korelacije med ocenjevalnim modelom in trenutno najboljšo možno izbiro pravil. Ta pravila so preverjena v praksi, saj jih uporabljajo pri prevajanju dnevnih izdaj časopisov. Pravila, ki jih sistem oceni kot neprimerna, lahko po pregledu strokovnjaka ustrezno popravimo oziroma izbrišemo.

  - Odkrivanje primernejših iskalnih algoritmov. V sistemu Apertium poteka izbira pravil po požrešni metodi algoritma najdaljšega možnega ujemanja z leve proti desni (left-to rigth longest match – LRLM rule selection). Preiskovanje vseh možnih pokritij izvorne povedi (iskanje optimalnega pokritja) je zamudno.
  - Ocenjevanje samodejno grajenih pravil. Metode za samodejno grajenje pravil (Sanchez-Martinez in Ney, 2006; Sanchez-Martinez in Forcada, 2007) izdelajo veliko število pravil; izbira najprimernejših pravil v določenih primerih je posebna domena ocenjevanja pravil.

- Raziskava algoritmov za izbiro pravil. V sistemu Apertium poteka izbira pravil po požrešni metodi algoritma najdaljšega možnega ujemanja iz leve proti desni (Left-to Rigth Longest Match Rule Selection – LRLM). V večini primerov se ta algoritem lepo sklada s človeškim načinom tvorjenja povedi, v določenih primerih pa ta algoritem ne najde najboljše rešitve, tj. najboljšega pokritja izvorne povedi. Odkrivanje takšnih primerov in iskanje boljšega algoritma predstavlja razdelek 7.2.4.1.
- Izdelava metrike za ocenjevanje pravil. V okviru doktorskega dela je bila izdelana posebna metrika za ocenjevanje pravil strukturnega transferja. Preizkušena je bila na ročno grajenih pravilih preizkušenega prevajalnega sistem (Forcada, 2006) in na nepreizkušenih pravilih novega testnega sistema (Vičič in Forcada, 2008).

5. **Hitra izdelava prevajalnega sistema na osnovi pravil plitkega transferja za sorodne jezike.** Osnova sistema je Apertium, ki je predstavljen v razdelku 3.4. Osnovni vodili snovanja sistema sta:

   - Omogočanje enostavnega dodajanja novih metod ter preizkušanja njihove uporabnosti, kar omogoča enostavno in kar najbolj objektivno evalvacijo opisanih prispevkov k znanosti. Metode so bile preizkušene na dejanski uporabi in ne le v umetno ustvarjenih okoljih.
   - Pri snovanju in postavljanju sistema so bili izdelani napotki za hitro izdelavo podobnih sistemov z drugimi jezikovnimi pari.

Poglavje 3

Sistemi za strojno prevajanje

Strojno prevajanje (Machine translation – MT) predstavlja vsako uporabo računalnikov kot pripomočkov za prevajanje besedil iz enega naravnega jezika v drugi (EAMT, 2010). V tem delu bomo obravnavali le strojno prevajanje naravnih jezikov brez uporabnikovega sodelovanja (fully automatic machine translation – FAMT, machine translation with no user intervention).

3.1 Razdelitev

Sodobni pregled strojnega prevajanja (Sanchez-Martnez et al., 2007) deli področje na dve skupini: prevajanje s pomočjo pravil (Rule-Based – RBMT) in prevajanje na osnovi korpusov (Corpus-Based – CBMT).

- RBMT obsega sisteme in metode za prevajanje s pomočjo zbirke pravil. Način zapisa pravil se med sistemi razlikuje, veže pa jih skupno dejstvo, da je postavitev takšnega sistema dolgotrajno opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, čeprav se pri gradnji uporabljajo nekateri manj standardni prijemi. Primeri sistemov: Systran[1], Promt[2], Apertium[3].

- CBMT obsega sisteme, ki sledijo naslednjemu vzoru: pripravljena je množica referenčnih prevodov, ki so analizirani in prevedeni v modele prevajalnega sistema po načelih, ki določajo prevajalni sistem (faza učenja). Ti modeli služijo kot osnova za poznejše prevode neznanih povedi (faza prevajanja). Sis-

[1]Systran: http://www.systran.co.uk/
[2]Promt: http://www.e-promt.com/
[3]Apertium: http://www.apertium.org/

teme te skupine delimo na dve večji podskupini: na sisteme statističnega strojnega prevajanja (Statistical Machine Translation – SMT (Al-Onaizan et al., 1999; Burbank et al., 2005)) in na sisteme strojnega prevajanja na osnovi primerov (Example Based Machine Translation – EBMT (Nagao, 1984)). Primeri sistemov: Google Translate (Och, 2006), Moses (Koehn et al., 2007), Egypt toolkit (EGYPT, 2007) in Genpar toolkit (GenPar, 2010).

- Hibridni sistemi predstavljajo mešanico obeh pristopov. Osnova takšnih sistemov sodi v eno od predstavljenih paradigem in je oplemenitena z metodami druge paradigme.

Ta delitev je več kot le teoretična, saj kar nekaj sistemov, ki se danes vsakodnevno uporabljajo, sodi v eno od obeh opisanih kategorij. Hibridni sistemi poskušajo z uporabo mešanih prijemov izboljšati kakovost oziroma odpraviti pomanjkljivosti osnovnih sistemov. Začetni prevajalni sistemi so bili postavljeni kot zbirke pravil, saj se je dostopnost elektronskih gradiv povečala šele v zadnjem času. Vseeno pa sistemov na osnovi pravil ne smemo zanemarjati, saj vsebujejo kar nekaj prednosti, kot sta natančna sledljivost prevajalnih postopkov in enostavno dopolnjevanje (Forcada, 2006). Sistemi, temelječi na metodah RBMT, dosegajo visoke rezultate tudi na račun visokih stroškov postavitve (Arnold, 2003). Sistemi, temelječi na metodah CBMT, kot so sistemi SMT (Brown et al., 1993) in EBMT (Nagao, 1984), omogočajo hitro postavitev prevajalnih sistemov ob predpostavki, da za izbrane jezikovne pare obstajajo veliki dvojezični korpusi, kar pa ni vedno res, predvsem za manj uporabljene jezike (Forcada, 2006).

Prevajanje poteka na več ravneh. Večina avtorjev tako predstavlja prevajalne sisteme kot skupek več modelov (Brown et al., 1993; Sanchez-Martnez et al., 2007; Burbank et al., 2005).

3.1.1 Statistično strojno prevajanje

Statistično strojno prevajanje (Statistical Machine Translation – SMT) je osnovano na parametričnih statističnih modelih, ki so naučeni na poravnanih dvojezičnih korpusih (učnih primerih).

Že od nekdaj je poskušal človek opisati jezik s pomočjo pravil. Prvi primeri segajo vsaj 2000 let nazaj. Pri opisovanju večine naravnih jezikov s strogimi pravili pa se pojavi kup problemov. Naravni jezik je preveč kompleksna in živa tvorba, pravila za opisovanje pa so preveč kompleksna, če jih je sploh mogoče vsa zapisati. Že v začetku tega stoletja so prišli strokovnjaki do tega zaključka: „All grammars leak" (vse gramatike puščajo) (Sapir, 1921).

Natančno določanje pravil jezika in umeščanje v stroge okvire pravil ni obrodilo sadov; potrebujemo namreč bolj ohlapne omejitve, ki pri uporabi jezika upoštevajo tudi ustvarjalnost.

Namesto analiziranja povedi po slovničnih pravilih iščemo splošne vzorce, ki se porajajo pri uporabi jezika. Glavno orodje za iskanje takšnih vzorcev je štetje raznovrstnih objektov, bolj strokovno izraženo, statistika. Od tod izvira tudi ime statistično strojno prevajanje.

3.1.2 Statistično strojno prevajanje z razčlenjevanjem

Statistično strojno prevajanje z razčlenjevanjem (statistical machine translation by parsing – SMTbyP) je posebna različica statističnega strojnega prevajanja, ki temelji na drevesnih strukturah in slovnicah. Snovalci sistemov statističnega strojnega prevajanja vedno pogosteje uporabljajo modele, temelječe na drevesnih strukturah in slovnicah (tree structured translation models) (Eng et al., 2003; Koehn et al., 2003).

Melamed (2004a) predlaga zmanjšanje konceptualne kompleksnosti prevajalnih modelov, temelječih na drevesih, in tudi novo ime za področje statistično strojno prevajanje z razčlenjevanjem (statistical machine translation by parsing – SMTbyP). GenPar (Burbank et al., 2005) je popoln sistem za postavitev prevajalnega sistema po načelih SMTbyP (Melamed, 2004a,b).

Prvi pogoj za sistem SMTbyP je vzporedni, povedno poravnan in skladenjsko označen dvojezični korpus in enojezični skladenjsko označen korpus (Melamed, 2004a). Primer skladenjsko označenega korpusa je Penn treebank (Marcus et al., 1993).

Osnovni sistem SMTbyP je sestavljen iz dveh faz: učne in prevajalne faze.

Prvi, učni del, uporablja skladenjski razčlenjevalnik, kot na primer v (Collins, 2003; Charniak, 2000), ki je bil vnaprej naučen na dvojezičnem skladenjsko označenem korpusu (Marcus et al., 1993). Vsaka poved iz izvornega in ciljnega dela korpusa je razčlenjena; rezultat so pari skladenjsko označenih, vzporednih (prevodov) izvornih in ciljnih povedi. Naslednji korak sestavlja hierarhične poravnave med drevesi izpeljave izvornih ter ciljnih povedi. Model statistične poravnave besed (Brown et al., 1993; Wu, 2005) je uporabljen za modeliranje povezav (prevodov) med besedami v korpusu. Učni podatki so shranjeni za poznejšo uporabo v prevajalnem delu.

V drugem, prevajalnem delu, sistem sestavi skladenjsko drevo vhodne povedi v izvornem jeziku (povedi, ki jo sistem prevaja). Na osnovi naučenih podatkov iz prve faze izdela primerno skladenjsko drevo v ciljnem jeziku in zamenja izvorne besede s ciljnimi s pomočjo modela statistične poravnave besed.

3.1.3 Strojno prevajanje na osnovi primerov

Strojno prevajanje na osnovi primerov (Example-based Machine Translation – EBMT) (Nagao, 1984) je pristop k strojnemu prevajanju, ki temelji na vzporednih dvojezičnih korpusih. V bistvu je prevajanje po analogiji. Sistemi EBMT razbijajo dele besedila na manjše enote in iščejo že poznane dele za prevod, te dele ponovno združujejo v končni izdelek. Če ta način primerjamo s prevajanjem človeka, naj bi prav ta princip najbliže odražal dejanski način prevajanja pri človeku. Prevajanje ne poteka z globoko jezikoslovno analizo izvornih besedil; besedila se razdelijo na manjše enote do te mere, da so posamezni kosi že poznani (examples), te dele prevedemo po analogiji (že videnem) in sestavimo končni izdelek.

Sistemi za strojno prevajanje na osnovi primerov pri učenju kodirajo abstrahirano znanje v obliki primerov s spremenljivkami.

3.1.4 Strojno prevajanje na osnovi pravil

Strojno prevajanje na osnovi pravil (Rule-Based Machine Translation – RBMT) obsega sisteme in metode za prevajanje s pomočjo zbirke pravil. Način zapisa pravil se med sistemi razlikuje, veže pa jih skupno dejstvo, da je postavitev takšnega sistema dolgotrajno opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, čeprav pri gradnji uporabljajo nekatere manj standardne prijeme. Primeri sistemov: Systran (Systran, 2010), Promt (Promt, 2010) in Apertium (Apertium, 2010).

Sistemi te paradigme izvorno besedilo najprej oblikoskladenjsko označijo in skladenjsko razčlenijo ter izdelajo predstavitev vhodnega besedila, po navadi v obliki skladenjskega drevesa izpeljave. Ta predstavitev se še dodatno abstrahira s poudarkom na zahtevah strojnega prevajanja. Proces prenosa prevede abstraktno predstavitev vhodnega besedila v izvornem jeziku v podobno predstavitev v ciljnem jeziku. To predstavitev sistem uporabi kot osnovo za tvorjenje besedila v ciljnem jeziku, v bistvu uporabi inverzne metode prvega dela na ciljnem jeziku.

3.1.5 Strojno prevajanje na osnovi pravil plitkega prenosa

Sistemi strojnega prevajanja na osnovi pravil plitkega razčlenjevanja in plitkega prenosa (shallow transfer rule based machine translation) v večini primerov uporabljajo enostavno arhitekturo, pri čemer je razčlenjevanje izvornega jezika omejeno na oblikoskladenjske oznake. Sistemi večinoma uporabljajo plitko razčlenjevanje (Homola in Kuboň, 2008a).

Večina sistemov za prevajanje sorodnih jezikov temelji na strojnem prevajanju s pravili plitkega razčlenjevanja, kot je pokazano v (Homola et al., 2009).

Metode popolne slovnične analize ne dosegajo dovolj dobrih rezultatov za uporabo v sistemih za prevajanje sorodnih jezikov, saj je njihova stopnja napak višja kot pa prednosti, ki jih omogoča tak način razčlenjevanja izvornih besedil.

3.1.6 Strojno prevajanje sorodnih naravnih jezikov v poljubnih domenah in nesorodnih naravnih jezikov v ozko omejenih domenah

Samodejno prevajanje naravnih jezikov visoke kakovosti (Fully Automatic High Quality Machine Translation) predstavlja hudo prepreko (EAMT, 2010), saj so jeziki spreminjajoče se tvorbe, ki jih je težko opisati. Izdelava modelov, ki dovolj dobro opisujejo prevajanje poljubnih besedil med poljubnimi jeziki, zahteva ogromna sredstva in ogromno časa. Probleme poskušamo omejevati s pomočjo poenostavljanj. Primeri takšnih poenostavitev prevajalnih problemov vključujejo:

- prevajanje s slabimi in nenatančnimi prevodi,
- prevajanje med sorodnimi jeziki,
- prevajanje nesorodnih jezikov v ozko omejeni domeni.

Prva možnost se zdi popolnoma neuporabna, vendar se takšni sistemi uporabljajo kot priročno, pogosto edino orodje za razumevanje gradiva. Njihova izdelava je enostavnejša. Obstaja kar nekaj kakovostnih orodij in metod, ki temeljijo na statističnem strojnem prevajanju (razdelek 3.1.1) in omogočajo enostavno postavitev takšnih prevajalnih sistemov. Primeri orodij so opisani v delih (Al-Onaizan et al., 1999; Burbank et al., 2005). Kakovost takšnih sistemov je zelo odvisna od velikosti učnih gradiv; velikost poravnanih dvojezičnih korpusov naj bi segala vsaj v desetine milijonov besed (Och, 2006). Večina sistemov, temelječih na teh metodah, je tako namenjena predvsem razumevanju besedila.

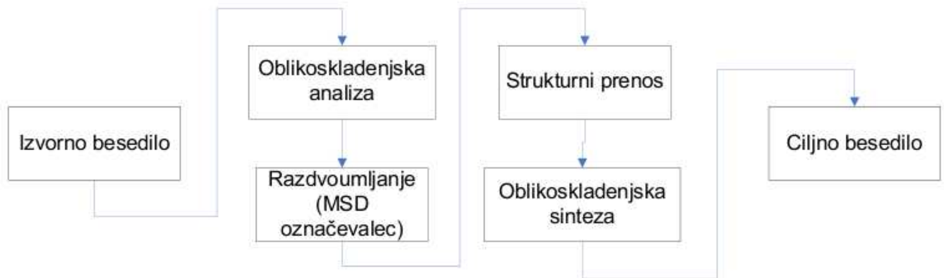
Prevajanje med sorodnimi jeziki predstavlja poenostavitev problema prevajanja tujih si jezikovnih parov oziroma popolnoma prostega prevajanja naravnih jezikov z omejitvijo razlik predvsem v strukturi povedi. Seveda pa podobnost ni omejena le na strukturo povedi, izraža se na vseh jezikovnih ravneh, tj. glasoslovju, besedoslovju, oblikoslovju, skladnji, (povzeto po (Toporišič, 2000)). Podobnosti na vseh naštetih ravneh lajšajo gradnjo prevajalnih sistemov.

Prevajanje med nesorodnimi jeziki v ozko omejeni domeni lahko v marsičem enačimo s prevajanjem sorodnih jezikov, saj ozko omejene domene po navadi prinašajo omejen besedni zaklad, omejen slog pisanja in podobno. Tako lahko mnoge metode, ki so zasnovane za prevajanje med sorodnimi jeziki, uporabimo tudi v primeru prevajanja med nesorodnimi jeziki v ozko omejenih domenah.

Ena od metod, ki omogoča relativno dobre rezultate prevodov sorodnih jezikov, je strojno prevajanje na osnovi pravil plitkega prenosa (shallow transfer – RBMT). Metoda ima že dolgo zgodovino in je bila uspešno uporabljena v mnogih prevajalnih sistemih, od katerih je najbolj znan Apertium (Corbi-Bellot et al., 2005).

3.2 Strojno prevajanje na osnovi pravil plitkega prenosa

Sistemi strojnega prevajanja s pravili plitkega prenosa (shallow transfer rule based machine translation) v večini primerov uporabljajo enostavno arhitekturo, pri čemer je analiza izvornega jezika omejena na oblikoskladenjske oznake. Arhitektura, ki jo uporablja večina sistemov za strojno prevajanje naravnih jezikov na osnovi pravil plitkega prenosa in plitke sinteze, je prikazana na sliki 3.1.



Slika 3.1: Moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Ta arhitektura je bila najprej predstavljena v (Hajič et al., 2000) in pozneje uporabljena tudi v (Corbi-Bellot et al., 2005).

Opis posameznih modulov prevajalnega sistema, kot so prikazani na sliki 3.1:

- Oblikoskladenjska analiza (morphosyntactic analysis) izvornega besedila vsaki besedi pripiše vse možne oblikoskladenjske oznake, ki bi jih ta besedna oblika lahko imela.
- Razdvoumljanje (disambiguation) služi za izbiro najverjetnejše oznake za posamezno besedo glede na njeno okolico.
- Strukturni prenos s pomočjo pravil in dobesednih prevodov prenese označeno besedilo v ciljni jezik.

- Oblikoskladenjska sinteza nadomesti oblikoskladenjsko označeno besedilo z dejanskimi besednimi oblikami v ciljnem jeziku.

Moduli so natančneje opisani v razdelku 3.4.1, in sicer na primeru ogrodja Apertium.

3.3 Orodja za postavitev prevajalnih sistemov

Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev samodejnega prevajalnega sistema za sorodne jezike. Obširneje je predstavljeno v razdelku 3.4, saj je bilo uporabljeno pri snovanju in preizkušanju večine metod, predstavljenih v tem delu. Apertium je licenciran pod licenco GNU Lesser General Public License (LGPL) (GNU, 2010).

GenPar (Burbank et al., 2005) je zbirka orodij za raziskave posplošenega razčlenjevanja (generalized parsing), predvsem strojnega prevajanja z razčlenjevanjem. Ta zbirka ponuja arhitekturo, načrt in implementacijo sistema za statistično strojno prevajanje z razčlenjevanjem (Statistical Machine Translation by Parsing SMTbyP).

Zbirka je prosto dostopna v okviru licence GPL (GNU, 2010), kar pomeni, da je poleg vseh programov prosto dostopna tudi izvorna koda. Pripravljena so že testna okolja s primeri učnih in testnih podatkov. Zbirka vsebuje vse programe za takojšnje preizkušanje sistemov ali implementacijo lastnih idej v že pripravljenem ogrodju. GenPar je licenciran pod licenco GNU General Public License (GPL) (GNU, 2010).

Egypt (Al-Onaizan et al., 1999) je zbirka orodij za postavitev sistema statističnega strojnega prevajanja. Na poletni delavnici leta 1999 na JHU (John Hopkins University) so po vzoru (Brown et al., 1994) izdelali zbirko orodij, ki omogočajo postavitev popolnega sistema za statistično strojno prevajanje, osnovanega na dvojezičnih vzporednih korpusih. Zbirko so poimenovali Egypt. Pri snovanju delavnice so si zadali pet osnovnih ciljev in jih tudi dosegli:

1. Postavitev zbirke orodij za statistično strojno prevajanje. Zbirka je dosegljiva vsej raziskovalni srenji. Sestavljena je iz orodij za pripravo korpusov, orodij za dvojezično učenje (postavitev parametričnih modelov) in orodij za takojšnje dekodiranje besedil.

2. Postavitev češko-angleškega sistema za prevajanje besedil na osnovi izdelanih orodij.

3. Osnovno testiranje sistema na osnovi objektivnih mer (statistično modeliranje težavnosti).

4. Izboljšanje osnovnih rezultatov z uporabo oblikoslovnih in skladenjskih prevajalnikov.

5. V zadnjih dneh delavnice so postavili prevajalni sistem za nek nov jezik v enem samem dnevu (potrditev enostavnosti uporabe orodij).

Moses (Koehn et al., 2007) je v zadnjem času najbolj uporabljano ogrodje za postavitev sistemov za statistično strojno prevajanje. Glavne lastnosti ogrodja so:

- dva tipa prevajalnih modelov (translation models): na osnovi fraz, dejansko delov besedila (phrase-based) in na osnovi dreves (tree-based);
- do določene mere omogoča integracijo eksplicitnega jezikovnega znanja na nivoju besed;
- omogoča podporo za integracijo orodij z dvoumnimi izhodi, kot so oblikoskladenjski analizatorji in razpoznavalniki govora;
- podpira velike jezikovne modele.

Moses je licenciran pod licenco GNU Lesser General Public License (LGPL) (GNU, 2010).

3.4 Apertium – odprtokodno ogrodje za prevajalni sistem sorodnih jezikov

Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev samodejnega prevajalnega sistema za sorodne jezike tipa plitkega prenosa (shallow transfer) (Sanchez-Martinez in Ney, 2006). Predstavlja ogrodje, ki omogoča s pomočjo pravil prevajanje med sorodnimi jeziki. Uvršča se med sisteme za samodejno prevajanje naravnih jezikov na osnovi pravil plitkega prenosa (shallow-transfer RBMT). Prevajanje je razdeljeno na pet osnovnih faz:

- označevanje neprevajanih razdelkov;
- leksikalni prenos;
- odpravljanje dvoumnosti (disambiguation);

- strukturni prenos;
- dejanski prevod posameznih besed in besednih zvez.

Zadnja faza odpravlja pomanjkljivosti ostalih faz in obsega upoštevanje niza pravil, ki pri prevajanju posebnosti odpravijo manjše napake.

Pravila, ki omogočajo prevajanje, temeljijo na regularnih jezikih, ki jih je enostavno pretvoriti v končne avtomate (transduktorje končnih stanj – finite state transducers), ter na besednih in fraznih slovarjih (enojezičnih in večjezičnih). Opisani sistemi so primerni le za prevajanje med sorodnimi jeziki, saj enostavna pravila ne omogočajo dobrega opisa poljubnih prevajalnih konstruktov. Več o tem je opisano v razdelkih 5.1 in 2.1.3.

Za leksikalni prenos so opisi s pomočjo regularnih jezikov oziroma iz teh izvedeni stohastični regularni modeli dovolj močno orodje (Melamed, 2004a). Primeri stohastičnih regularnih modelov so: Skriti Markovski Model (hidden markov model – HMM), predstavljen v (Welch, 2003), ali uteženi končni transduktorji (weighed finite state transducers – WFST), definirani v (Kornai, 1999; Roche in Schabes, 1997).

Strukturni nivo je kompleksnejši, saj nekaterih konstruktov oziroma njihovih prevodov ne moremo opisati z regularnimi jeziki. Mnogi avtorji (Melamed, 2004b; Eng et al., 2003; Koehn et al., 2003) predlagajo uporabo modelov, ki temeljijo vsaj na drevesnih strukturah. Tako je Apertium namenjen predvsem za sorodne jezike.

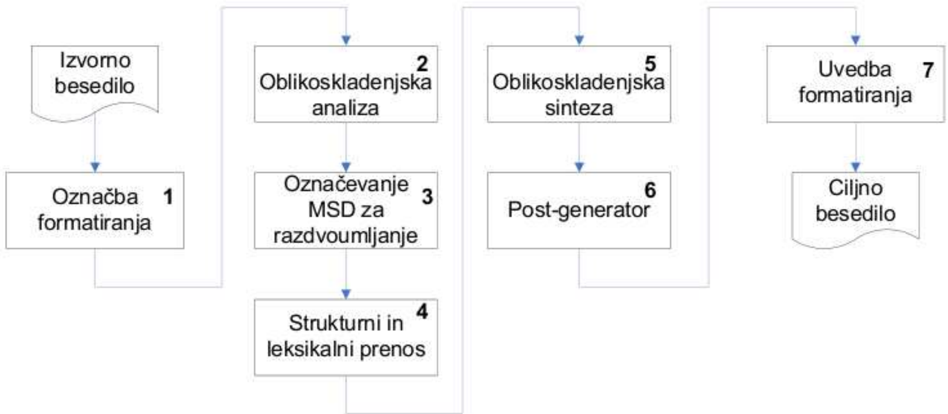
3.4.1 Arhitektura ogrodja Apertium

Arhitektura ogrodja Apertium posnema arhitekturo, predstavljeno na sliki 3.1, ki kaže tipično razporeditev modulov sistemov za strojno prevajanje s plitkim prenosom.

Slika 3.2 predstavlja arhitekturo ogrodja Apertium.

Prvi modul izbere le besedilo, ki ga prevajamo iz označenega besedila, kar pomeni, da za ostale prevajalne module izbriše oznake formatiranja, kot so oznake jezika HTML, posebej označi dele besedila, ki se ne prevaja, na primer za vnaprej pripravljene obrazce. Delov besedila, ki se ne prevajajo, pa ne izbriše, ampak jih posebej označi, tako da jih ostali moduli prevajalnega sistema ignorirajo.

Oblikoskladenjska analiza izvornega besedila poišče vse možne kombinacije oblikoskladenjskih oznak za posamezne besede glede na pripravljene leksikone izvornega jezika. Razdvoumljanje izbere najverjetnejše oznake za posamezne besede glede na njeno okolico, strukturni prenos pa služi dejanskemu prenosu izvornega besedila v ciljni jezik. Modul postgenerator služi za odpravljanje napak, uvajanje posebnosti ciljnega jezika in za združevanje besednih zvez. Oblikoskladenjska sinteza poišče ustrezne besede v pripravljenem leksikonu ciljnega jezika glede na prevedene oblikoskladenjsko označene dele besedila. Sledi ponovno dodajanje delov besedila, ki ga je prvi modul posebej označil in se zato niso prevajali. Moduli so natančneje predstavljeni v razdelku 3.4.2.



Slika 3.2: Arhitektura ogrodja Apertium: poleg osnovnih modulov, ki služijo za osnovno prevajanje in so prikazani na sliki 3.1, Apertium dodaja še module za označevanje delov besedila, ki se ne prevajajo, in modul za končno urejanje (post-editing) prevodov.

Sledijo opisi posameznih modulov prevajalnega sistema, kot so prikazani na sliki 3.2; moduli so označeni s števili.

Modul 1: označba formatiranja (De-formatter). V izvornem besedilu posebej označi dele besedila, ki jih ostali prevajalni moduli ignorirajo. Tako lahko sistem prevaja tudi besedilo z urejevalnimi oznakami, kot so oznake jezikov HTML ali XML. Za posebej prirejene sisteme lahko modul označuje tudi dele besedila, ki se ne prevajajo.

Modul 2: oblikoskladenjska analiza (Morphological analyzer). Vsaki besedi izvornega besedila pripiše vse možne oblikoskladenjske oznake, ki bi jih ta besedna oblika lahko imela. Modul podpira besede in besedne zveze, ki so zapisane v oblikoskladenjsko označenem enojezičnem slovarju izvornega jezika. Vsaka beseda oziroma besedna zveza je obdelana samostojno, brez vpliva okolice. Z izvajanjem oblikoskladenjske analize vsaki besedi pripiše vse njene možne oznake, kar pomeni, da je izhod tega modula dvoumen. Primer delovanja je prikazan na sliki 3.3. V pomoč pri sledenju služi tabela 3.1, v kateri so predstavljene kratice oznak MSD po specifikacijah JOS (Erjavec et al., 2009).

```
Danes je lepo vreme
Danes
    danes Rsn
je
    biti  Gp-ste-n
    jesti Ggnspm
lepo
    lep   Ppnsei
    lep   Ppnset
    lep   Ppnzeo
    lep   Ppnzet
vreme
    vreme Sosei
    vreme Soset
.
```

Slika 3.3: Oblikoskladenjska analiza stavka „Danes je lepo vreme”. Besede izvorne povedi so označene z vsemi možnimi ustreznimi oblikoskladenjskimi oznakami iz slovarja. Najprej je zapisana besedna oblika, sledijo vse možne oznake zanjo. Za besedno obliko *lepo* je možnih pet različnih množic oznak.

Modul 3: razdvoumljanje s označevanjem MSD (POS tagger). Služi za izbiro najverjetnejše oznake za posamezno besedo glede na njeno okolico. Sistem uporablja stohastični označevalnik oblikoskladenjskih oznak, ki iz seznama možnih oblikoskladenjskih oznak, ki jih je našel modul za oblikoskladenjsko analizo, izbere najverjetnejšo oznako.

Osnovni označevalnik (Sánchez-Martínez et al., 2008) temelji na samodejni metodi učenja označevanja oblikoskladenjskih oznak na neoznačenem korpusu. Po trditvah avtorjev je primeren za vse evropske jezike, kakovost označevanja pa ne dosega točnosti najboljših označevalcev oziroma označevalcev, naučenih na označenih in pregledanih korpusih. Rezultat tega modula je po ena izbira oblikoskladenjskega označevalca za vsako besedo izvornega besedila.

Modul 4: strukturni prenos (Structural transfer). S pomočjo pravil in dobesednih prevodov prenese označeno besedilo v ciljni jezik. Pravila strukturnega prenosa temeljijo na regularnih izrazih in se osredotočajo na ujemanja med oblikoskladenjskimi oznakami sosednih besed, na lokalnem besednem vrstnem redu in

Tabela 3.1: Razširjene kratice, ki so uporabljene v na sliki 3.3 in primerih 3.1, 3.2 in 3.3.

Rsn	prislov splošni nedoločno
Gp-ste-n	glagol pomožni sedanjik tretja ednina nikalnost (ne)
Ggnspm	glagol glavni nedovršni sedanjik prva množina
Ppnsei	pridevnik splošni nedoločno srednji ednina imenovalnik
Ppnset	pridevnik splošni nedoločno srednji ednina tožilnik
Ppnzeo	pridevnik splošni nedoločno ženski ednina orodnik
Ppnzet	pridevnik splošni nedoločno ženski ednina tožilnik
Sosei	samostalnik občno_ime srednji ednina imenovalnik
Soset	samostalnik občno_ime srednji ednina tožilnik
Pdnset	pridevnik deležniški nedoločno srednji ednina tožilnik
Soset	samostalnik občno_ime srednji ednina tožilnik
Pdnmet	pridevnik deležniški nedoločno moški ednina tožilnik
Somet	samostalnik občno_ime moški ednina tožilnik

na ostalih razlikah jezikovnega para, ki jih lahko opišemo z lokalnim kontekstom. Obširneje so pravila predstavljena v 5. poglavju.

Dobesedni prevodi posameznih besed v lematizirani obliki temeljijo na osnovi dvojezičnega slovarja.

Modul 5: oblikoskladenjska sinteza (Morphological analyzer). Oblikoskladenjsko označeno besedilo, ki je že prevedeno v ciljni jezik, nadomesti z dejanskimi besednimi oblikami v ciljnem jeziku. Sam modul uporablja enake podatkovne strukture (enojezični oblikoskladenjsko označen slovar) kot Modul 2, tj. modul za oblikoskladenjsko analizo, zamenjana je le smer uporabe.

Modul 6: končno urejanje (Post-generator). Služi za odpravljanje sistemskih napak prejšnjih modulov, za uvajanje posebnosti ciljnega jezika (uporaba diakritikov, posebnosti pri imenih ...) in za združevanje besednih zvez. Modul kot osnovno jezikovno gradivo uporablja slovar, ki pa ni oblikoskladenjsko označen. Primer 2.2 kaže združevanje določnega člena in predloga v italijanščini v predložno zvezo (preposizione articolata).

Slika 3.4 kaže del paradigme, ki združuje določni člen *la* in predlog *di* v predložno zvezo. Prikazana sta primera za ženski spol, ločeno za besede, ki se začnejo s samoglasnikom in soglasnikom. Ime paradigme je *di*.

```
<pardef n="di">
  <e>
    <p><l><b/><a/>la<b/></l><r>della<b/></r></p>
    <par n="soglasnik"/>
  </e>
  <e>
    <p><l><b/><a/>la<b/></l><r>dell'</r></p>
    <par n="samoglasnik"/>
  </e>
  ...
</pardef>
```

Slika 3.4: Del paradigme *di*, ki združuje določni člen *la* in predlog *di* v predložno zvezo.

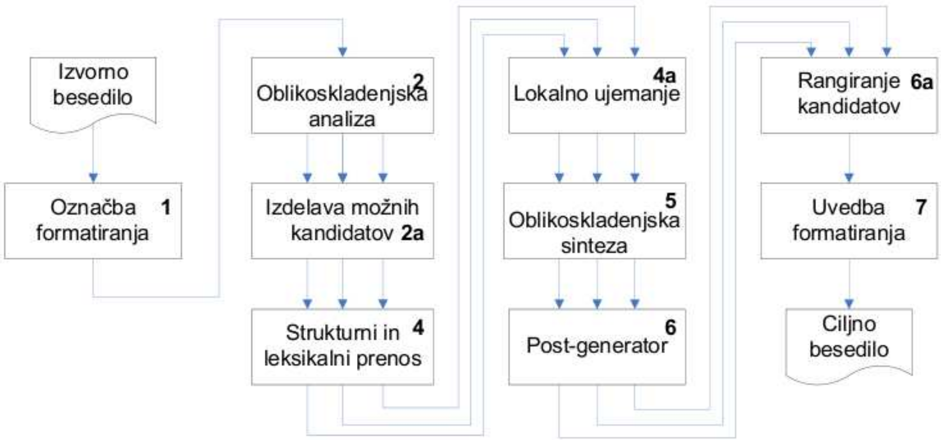
Modul 7: ponovna uvedba formatiranja (Re-formatter). Kot zadnji modul prevajalne verige ponovno postavi v besedilo odseke besedila, ki jih je začetni modul izbral in označil kot besedilo, ki se ne prevaja; največkrat so to oznake urejanja besedila (HTML, XML in ostale).

3.4.2 Predlagana spremenjena arhitektura

Rezultati, predstavljeni v (Homola in Kuboň, 2008b), kažejo, da sprememba arhitekture brez uporabe označevalnika MSD v začetnih fazah prevajanja in z uvedbo stohastičnega razvrščevalnika prevodov (stochastic ranker) na koncu prevajalne verige prinaša izboljšano kakovost prevodov v primerjavi z osnovno arhitekturo, predstavljeno na sliki 3.1 in v (Corbi-Bellot et al., 2005). Odločili smo se za izbiro spremenjene arhitekture, kot je prikazana na sliki 3.5, in jo predstavili v (Vičič et al., 2009).

Najpomembnejši razlogi za izbiro spremenjene arhitekture so:

- Izdelava označevalnika MSD, zlasti dobre kakovosti, ni enostavna naloga. Poseben problem predstavljajo oblikoslovno bogati jeziki, med katere sodi tudi slovenščina. Eden izmed najlažjih načinov je učenje stohastičnih označevalcev, ki temelji na algoritmu HMM (Welch, 2003). Nekatere dele te naloge lahko avtomatiziramo z uporabo nenadzorovanih ali delno nadzorovanih učnih metod, kot je (Brants, 2000), vendar še vedno ostaja dovolj dela z izbiro novega niza oznak, izdelavo označenega učnega korpusa, s preizkušanjem



Slika 3.5: Moduli predlaganega (spremenjenega) sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Arhitektura temelji na sistemu, predlaganem v (Corbi-Bellot et al., 2005; Hajič et al., 2003), brez uporabe sistema za razdvoumljanje na osnovi označevalnika MSD z uporabo vseh kandidatov za prevode do zadnjih faz prevajalne verige ter z dodatkom modula za izbiro najboljšega prevoda (Ranker).

korpusa in na koncu z izvedbo samega učnega procesa.

- Stopnja kakovosti označevanja današnjih najboljših (state-of-the-art) označevalnikov MSD za visoko pregibne jezike, kot sta opisana v (Hajič, 2000; Erjavec et al., 2005), je relativno nizka v primerjavi s kakovostjo označevalnikov MSD za analitične jezike, kot je angleški jezik, in tudi v primerjavi s splošno kakovostjo prevajalnih sistemov za sorodne jezike.

- Po arhitekturi, predstavljeni na sliki 3.1, modul za oblikoskladenjsko razdvoumljanje sledi modulu za oblikoskladenjsko analizo izvornega jezika. Napake stohastičnih metod modula za razdvoumljanje, ki uporablja označevalnik MSD, so tako povzročene v zgodnjih fazah prevoda in povzročajo več težav kot napake poznejših faz prevajalnega procesa.

- V zaključni fazi, ko so zbrani vsi razpoložljivi podatki za prevod, omogoča večje število kandidatov boljšo izbiro.

Tako spremenjena arhitektura lahko povzroči kombinatorično eksplozijo možnih kandidatov za prevode, saj v prvi fazi ohranjamo vse dvoumnosti in število kandidatov dobimo kot produkt vseh dvoumnosti. Kombinatorično eksplozijo števila možnih kandidatov nadzorujemo s pravili za ujemanje med oblikoskladenjskimi oznakami sosednih besed izvornega jezika in z rangiranjem kandidatov za prevode. Izdelava kandidatov in omejevanje njihovega števila sta obširneje prikazana v naslednjem odstavku.

Sledi opis dodatnih modulov prevajalnega sistema, kot so prikazani na sliki 3.5. Opisani so le novi moduli; originalni moduli so namreč opisani v razdelku 3.4. Moduli so označeni s števili.

Modul 2a: izdelava možnih kandidatov (Multiple candidate selector). Ta modul predstavlja zamenjavo za modul za razdvoumljanje, saj namesto izbire najverjetnejšega kandidata za prevode izdela vse možne kandidate. Kandidate sestavi na osnovi dvoumnega označevanja oblikoskladenjskega analizatorja.

Modul za oblikoskladenjsko analizo poišče vse možne oznake za vhodno poved in za vsako besedo izdela množico dvoumnih oznak. Množica vseh možnih kandidatov za prevode se sestavi kot kartezični produkt množic dvoumnih oznak. Število kandidatov narašča eksponentno z dolžino povedi, osnova eksponenta pa s številom dvoumnosti. Enačba 3.1 ponazarja eksponentno naraščanje števila možnih kandidatov, v kateri je kp množica kandidatov za prevode za najdaljšo poved, S_{max} najdaljša poved in x_{max} največje število dvoumnosti za besedo iz jezika.

$$|kp| = \prod_{i=0}^{|S_{max}|} x_{max} \tag{3.1}$$

Takšno število možnih kandidatov bi dobili ob najdaljši možni povedi, ki bi bila sestavljena le iz besed z največjim številom dvoumnosti (Vičič in Homola, 2010). Takšne povedi v jeziku sicer ne obstajajo in ta vrednost predstavlja nedosegljivo (pesimistično) zgornjo mejo.

Čeprav enačba 3.2, ki pri izdelavi kandidatov za prevode upošteva povprečne vrednosti števila dvoumnosti, kaže veliko nižje številke, ostaja kompleksnost problema še vedno eksponentna.

$$|kp| = \prod_{i=0}^{|\bar{S}|} \bar{x} \tag{3.2}$$

Enačba 3.3 kaže izračun enačbe 3.1 z dejanskimi vrednostmi iz korpusa „1984”; največja množica dvoumnosti obsega 15 elementov. Enačba 3.4 kaže empirične vrednosti povprečja 200 primerov povedi iz korpusa „1984”; povprečno število dvoumnosti je bilo 2,5 in povprečna dolžina povedi 15.

$$
\begin{aligned}
|S| &= 40 \\
x &= 15 \\
|kp| &= \prod_{i=0}^{40} 15 = 110,573323209e+45
\end{aligned}
\tag{3.3}
$$

$$
\begin{aligned}
|\bar{S}| &= 15 \\
\bar{x} &= 2,5 \\
|kp| &= \prod_{i=0}^{15} 2,5 = 9.313.221
\end{aligned}
\tag{3.4}
$$

Število kandidatov za prevode zmanjšamo z uporabo pravil za lokalno ujemanje oblikoskladenjskih oznak sosednih besed. Gradnja teh pravil je predstavljena v razdelku 5.5.

Algoritem aplicira vsa pravila, ki se skladajo s kandidatom, vsa pravila, katerih regularni izraz opisuje del kandidata za prevod. Če vsaj eno pravilo spremeni kandidata, tega završemo. Metoda je natančneje predstavljena v (Vičič in Homola, 2010). Izhod modula je množica najboljših n kandidatov za prevode (n-best set).

Modul 4a: lokalno ujemanje (Local agreement). Modul skrbi za ujemanje bližnjih besed v oblikoskladenjskih oznakah. Modul s pomočjo pravil za lokalno ujemanje odpravlja napake, ki jih lahko povzročijo pravila za strukturni prenos, še posebej samodejno zgrajena pravila. Primer 3.1 kaže lokalno ujemanje pridevnika in samostalnika v slovenščini; *odprto okno*: pridevnik *odprto* se ujema s samostalnikom *okno* v spolu, številu in sklonu.

Pri prevodu v srbščino se spol samostalnika spremeni iz srednjega v moški *okno – prozor*. Primer 3.2 kaže napačen prevod, ki bi ga dobili brez pravil za lokalno ujemanje: *otvoreno prozor*, pridevnik in samostalnik se ne ujemata več v spolu. Primer 3.3 kaže točen prevod, pri katerem pravilo pridevniku spremeni spol: *otvoren prozor*, okrajšave oznak MSD so zapisane v razširjeni obliki v tabeli 3.1.

(3.1) *odprto okno*
Pdnset Soset
"odprto okno"

(3.2) *otvoren prozor*
Pdnset Somet
"otvoreno prozor"

(3.3) *otvoren prozor*
Pdnmet Somet
"otvoren prozor"

Modul 6a: rangiranje kandidatov (Stochastic ranker). Iz množice prevedenih kandidatov za prevode izbere najverjetnejši prevod. Osnova za ugotavljanje kakovosti prevoda je statistični jezikovni model ciljnega jezika. Uporabljen je trigramski statistični jezikovni model.

Bistveni del celotnega prevajalnega sistema je statistični modul za rangiranje prevodov (Ranker) (Vičič et al., 2009). Glavni problem spremenjene arhitekture je, da oblikoskladenjski analizator in strukturni prenos povzročata lokalne oblikoskladenjske in strukturne dvoumnosti (ambiguities). Njihova kombinacija nato v procesu prevajanja ustvari veliko število različic (hipoteze). Zelo bi bilo zapleteno, če sploh mogoče, ročno izdelati pravila, ki bi omogočila reševanje tovrstnih dvoumnosti. Zato je v osnovno arhitekturo dodan modul za rangiranje prevodov na osnovi stohastičnih metod, katerega cilj je izbira povedi, ki najbolj ustreza modelu ciljnega jezika; torej povedi, ki je slovnično najbolj pravilno zapisana v ciljnem jeziku. Takšno rangiranje prevodov nam samo zase ne zagotavlja, da bo izbrana poved najboljši prevod izvorne povedi; modul v ciljnem jeziku poišče le slovnično najbolj pravilno poved. Arhitektura se pri ohranjanju pomena v končnem prevodu zanaša na predhodne module.

Uporabljeni statistični jezikovni model sodi med modele na osnovi trigramov (Statistical Trigram Language Model). Povedi sestavimo v verige trigramov, treh zaporednih besed, verjetnost za vsako poved pa izračunamo po enačbi 2.1, torej zmnožimo verjetnosti vseh trigramov, ki sestavljajo poved. Verjetnost za vsak trigram izračunamo po enačbi 2.2, v kateri preštejemo število pojavitev trigrama in jo utežimo s številom pojavitev bi- in monogramov, ki sestavljajo ta trigram. Števila pojavitev mono-, bi- in trigramov preštejemo v učnem korpusu; uporabili smo lasten korpus, sestavljen iz člankov Wikipedie; korpus je obširneje predstavljen v razdelku 2.4.3.

Za izdelavo trigramskega statističnega modela jezika je bila oblikovana lastna programska oprema, rezultate pa smo primerjali z orodjem za statistično modeliranje (CMU – statistical language modelling toolkit) (Clarkson in Rosenfeld, 1997). Za 100 naključno izbranih povedi smo dobili enake rezultate, kar ne preseneča, saj so bile uporabljene enake metode. Izdelava lastne programske opreme omogoča večjo fleksibilnost pri snovanju sistema in lažje preizkušanje morebitnih izboljšav.

3.4.3 Pregled uporabljenih jezikovnih virov

Sledi spisek jezikovnih virov, ki jih potrebujemo za vse module spremenjenega prevajalnega sistema (novi jezikovni viri, ki jih zahteva spremenjena arhitektura, so označeni z zvezdico – $*$).

Enojezični slovar izvornega jezika z oblikoskladenjskimi oznakami se uporablja za oblikoskladenjsko analizo izvornega besedila. Enojezični slovar izvornega jezika je uporabljen pri oblikoskladenjski analizi besedil v modulu za oblikoskladenjsko analizo (morphological analyser module), označenem s številko 2 na slikah 3.2 in 3.5.

Enojezični slovar ciljnega jezika z oblikoskladenjskimi oznakami se uporablja za oblikoskladenjsko sintezo ciljnega besedila. Enojezični slovar ciljnega jezika je uporabljen pri sintezi prevedenega besedila v ciljni jezik v modulu za oblikoskladenjsko sintezo (morphological generator module), označenem s številko 5 na slikah 3.2 in 3.5. Ta dva modula uporabljata iste slovarje, kar omogoča hitro postavitev dvosmernega prevajalnega sistema.

Dvojezični prevajalni slovar se uporablja za dobesedno prevajanje in prevajanje fraz v lematizirani obliki. Modul, ki uporablja dvojezični slovar in pravila plitkega prenosa, je modul za strukturni prenos (structural transfer module), označen s številko 4 na slikah 3.2 in 3.5.

Pravila plitkega prenosa na osnovi regularnih izrazov so uporabljena za opis oblikoskladenjskih in skladenjskih razlik lokalnega obsega med jezikoma prevajanega jezikovnega para, kot so lokalno ujemanje besed v oblikoskladenjskih kategorijah in lokalni vrstni red besed. Natančneje so pravila predstavljena v 5. poglavju.

Pravila na osnovi regularnih izrazov za končno urejanje (postediting) uporablja modul za končno urejanje (postgenerator), označen s številko 6 na sliki 3.5. Pravila služijo za odpravljanje napak, uvajanje posebnosti ciljnega jezika in za združevanje besednih zvez.

$*$ **Pravila na osnovi regularnih izrazov za izražanje lokalnega ujemanja oblikoskladenjskih kategorij izvornega jezika** uporablja modul za izbiro množice kandidatov za prevod (multiple candidate selector), ki je na sliki 3.5 označen s številko 2a. Modul uporablja isto tehnologijo kot modul za strukturni prenos (Structural transfer module) s pravili za lokalno ujemanje oblikoskladenjskih kategorij, ki so bila naučena na izvornem jeziku. To metodo uporablja modul kot hevristiko za omejevanje eksplozije števila možnih hipotez za prevode dvoumne oblikoskladenjske analize izvornega besedila.

∗ **Pravila na osnovi regularnih izrazov za izražanje lokalnega ujemanja oblikoskladenjskih kategorij ciljnega jezika** so uporabljena v modulu za iskanje lokalnega ujemanja (local agreement module), ki je označen s številko 4a na sliki 3.5. Uporabljajo se za odpravo napak, ki jih povzročajo pravila plitkega prenosa v fazi prenosa. Delovanje tega modula je v osnovi enako delovanju modula za strukturni prenos; drugačna so le pravila, ki popravljajo napake lokalnega ujemanja oblikoskladenjskih kategorij, delovanje modula pa je preseljeno na izhod modula za strukturni prenos.

∗ **Statistični jezikovni model ciljnega jezika** se uporablja v modulu za stohastično rangiranje končnih prevodov (Ranker module), ki predstavlja razširitev osnovne arhitekture po (Hajič et al., 2003) in je bil opisan v (Homola in Kuboň, 2008b), označen s številko 6a na sliki 3.5. Ta modul izbere najboljši prevod, oziroma v ciljnem jeziku najverjetnejši prevod, in sicer iz seznama možnih kandidatov za prevode, ki so jih sestavili prejšnji moduli. Deluje na principu stohastičnega jezikovnega modela ciljnega jezika. Natančneje je razložen v razdelku 2.1.8.

∗ **Jezikovni model oblikoskladenjskih oznak izvornega jezika** uporablja modul za izbiro množice kandidatov za prevod (multiple candidate selector), označen pa je s številko 2a na sliki 3.5. Poleg predstavljene hevristike uporablja še enako tehnologijo kot modul za rangiranje prevodov, vendar s pomočjo jezikovnega modela, naučenega na oblikoskladenjskih oznakah korpusa v izvornem jeziku. Metoda je natančneje predstavljena v (Homola et al., 2009).

Zgornji spisek vsebuje vsa jezikovna gradiva, ki so potrebna za postavitev prevajalnih sistemov spremenjene arhitekture, kot je opisana v razdelku 3.4.2. Za vsako jezikovno gradivo smo poiskali metodo, ki omogoča samodejno izdelavo tega gradiva. Če metode nismo našli, oziroma še ni bila predstavljena ali pa so bila gradiva neprimerna, smo izdelali novo metodo. Metode so natančneje predstavljene v razdelku 4.3.

S pomočjo opisanih metod smo izdelali več popolnoma delujočih sistemov. Sistemi so predstavljeni v razdelku 7.2.1, kakovost prevodov pa je ovrednotena v 7. poglavju.

Poglavje 4

Leksikoni z oblikoskladenjskimi informacijami

Sistemi za strojno prevajanje na osnovi pravil plitkega prenosa uporabljajo pri izdelavi prevodov iz izvornega v ciljni jezik enostavne metode analize, prenosa in sinteze. Sama analiza izvornih povedi temelji na oblikoskladenjskem označevanju, ki se izvaja s pomočjo oblikoskladenjsko označenih enojezičnih slovarjev. Transfer uporablja pravila plitkega prenosa in dvojezični slovar izvornega in ciljnega jezika. Pri sintezi je uporabljen enojezični slovar ciljnega jezika, ki je sestavljen na enak način kot slovar, uporabljen pri analizi, le da se ga uporablja za sintezo besed in ne za označevanje izvornih besed.

4.1 Oblikoskladenjski slovar

Besedna oblika v slovenščini, tudi v večini ostalih indoevropskih jezikov, je sestavljena iz krna in končnice ter redkeje predpone. Besedne oblike z istim pomenom družimo v razrede – lekseme. Tipični predstavnik takega razreda je lema – kanonična oblika. Oblikoskladenjski slovar tako združuje vse besedne vrste v razrede s tipičnimi predstavniki – lemami. Nadalje te razrede, oziroma njihove tipične predstavnike druži v paradigme, razrede ki združujejo vse vse leme, ki se spreminjajo po istih pravilih glede na oblikoskladenjske oznake. Definicija **??** vsebuje formalno definicijo oblikoskladenjskega slovarja.

definicija 4.1.1. *Beseda w je sestavljena iz predpone pr, krna k in pripone po: $w = pr \circ k \circ po$. $\circ$ označuje stik dveh nizov.*

definicija 4.1.2. *Oblikoskladenjska oznaka (morphosyntactic description) MSD vsaki besedni obliki pripiše njen razred. Vsaki besedi w pripada oblikoskladenjska oznaka MSD, torej beseda pripada množici, ki jo opisuje ta oznaka: $w \in MSD_i$.*

definicija 4.1.3. *Leksem Le je sestavljen iz besed z istim pomenom, oblika besed je definirana z množico pravil Le_p. Množico pravil Le_p sestavljajo pravila oblike $p_i \in Le_p; p_i \equiv pr_i \circ k_i \circ po_i \Rightarrow MSD_i$.*

definicija 4.1.4. *Tipični predstavnik leksema Le je beseda v kanonski obliki, imenujemo jo lema lm.*

definicija 4.1.5. *Beseda w pripada določenemu leksemu Le, če obstaja vsaj eno pravilo $p_i \in Le_p$, za katero velja:*
$w \in Le \Rightarrow \exists p_i \in Le_p; w = pr \circ k \circ po; p_i \equiv pr_i \circ k_i \circ po_i \Rightarrow MSD_i; pr_i = pr; k_i = k; po_i = po; w \in MSD_i$.

definicija 4.1.6. *Paradigma P združuje lekseme, ki vsebujejo nabor pravil, ki se razlikujejo le po krnih: $Le \in P \Rightarrow \forall p_i \equiv pr_i \circ k_i \circ po_i \Rightarrow MSD_i \in Le_i \in P \exists p_j \equiv pr_j \circ k_j \circ po_j \Rightarrow MSD_j \in Le : pr_i = pr_j; po_i = po_j; MSD_i = MSD_j;$*

definicija 4.1.7. *Oblikoskladenjski slovar sestavljajo paradigme iz definicije 4.1.6 ter leme iz definicije 4.1.4 s povezavo na paradigme: $lm \in P_i$.*

Oblikoskladenjski slovar, ki ga uporablja Apertium – lahko pa bi takšne slovarje z manjšimi spremembami uporabljali tudi drugi prevajalni sistemi – temelji na lemah, ki so zbrane v paradigmah. Posamezna paradigma združuje vse leme, ki se spreminjajo po istih pravilih glede na oblikoskladenjske oznake.

```
<e lm="cepljen"><i>cepljen</i><par n="žveplen/__pridevnik"/></e>
<e lm="procesija"><i>procesij</i><par n="žog/a__samostalnik"/></e>
<e lm="cerkev"><i>cerk</i><par n="cerk/ev__samostalnik"/></e>

lema: cerkev
krn: cerk
paradigma: cerk/ev__samostalnik
```

Slika 4.1: Del zapisov v enojezičnem slovarju. Lema je zapisana v atributu *lm* značke *e*, nato sledi krn ter značka *par*, ki označuje paradigmo. Zapis *cerkev* je predstavljen z lemo, krnom ter paradigmo. Značke so obširneje predstavljene v tabeli 4.1.

Slika 4.1 kaže primere lem in njihovo članstvo v paradigmah. Lema je predstavljena s svojim imenom (ime leme), krnom, najdaljšim delom, ki je skupen vsem

Tabela 4.1: Razlaga značk in atributov zapisa slovarjev v formatu Apertium.

značka	opis
<pardef>	definicija paradigme
<e>	element for entry – zapis v slovarju in paradigmi
<p>	string pair – par nizov
<par>	reference to paradigm – povezava na paradigmo
<re>	reference to regular expression – povezava na regularni izraz
<s>	reference to regular symbol – povezava na simbole oblikoskladenjskih oznak
<i>	reference to inentity transduction – način za zapis para nizov z isto vsebino
<l>	left part – leva stran zapisa besedila s slovničnimi simboli
<r>	right part – desna stran zapisa besedila s slovničnimi simboli
<lm>	lema
atribut	**opis**
n	dejanska vsebina značke <s>

besednim oblikam leme, in z imenom paradigme, v kateri so opisana vsa pravila sprememb glede na oblikoskladenjske kategorije. Posamezen zapis v slovarju je predstavljen z oznako XML *e*, atribut te oznake *lm* predstavlja ime leme, gnezdena oznaka *i* krn besede, oznaka *par* pa ime paradigme.

Primer za lemo *cerkev* je predstavljen na sliki 4.1. Posamezna gesla enojezičnega slovarja so združena v oblikoskladenjske paradigme, kot so definirane v (Spencer, 1991). Oblikoskladenjske paradigme predstavljajo razrede lem, ki se spreminjajo na isti način (glede na vse možne besedne oblike). Z drugimi besedami: vsebujejo vse leme, katerih vse besedne oblike se spreminjajo na enak način za vse oblikoskladenjske oznake MSD, ki so obširneje predstavljene v razdelku 2.1.1.2.

Slika 4.2 kaže primer paradigme za ženski samostalnik v slovenščini.

Uporaba paradigme omogoča izdelavo kompaktnejšega zapisa podatkov, kot je prikazano na sliki 4.2. Za paradigmo *cerk/ev__samostalnik* v slovenskem jeziku velja: vsi samostalniki prve ženske sklanjatve paradigme *-ev*, kot so cerkev, breskev, podkev, se sklanjajo po istem vzorcu in jih združimo v isto paradigmo. Enostavno pravilo določa spremembo besede iz imenovalnika v rodilnik s spremembo končnice iz cerkev v cerkve, torej $-ev \leftarrow -ve$. Eno pravilo tako zadošča za celo skupino besed in ne le za en osamljen primer.

```
<pardef n="cerk/ev__samostalnik">
  <e>
    <p>
      <l>
        ev
      </l>
      <r>
        ve
        <s n="samostalnik"/>
        <s n="ženski"/>
        <s n="ednina"/>
        <s n="imenovalnik"/>
      </r>
    </p>
  </e>
  <e>
    <p>
      <l>
        ev
      </l>
      <r>
        ev
        <s n="samostalnik"/>
        <s n="ženski"/>
        <s n="ednina"/>
        <s n="rodilnik"/>
      </r>
    </p>
  </e>
  ...
</pardef>
```

Slika 4.2: Del paradigme za samostalnike ženskega spola v slovenščini. Tipični predstavnik je lema *cerkev*. Končnica *-ev* se spreminja v skladu z različnimi MSD. Značke so obširneje predstavljene v tabeli 4.1.

Pri indoevropskih jezikih, ki večinoma uporabljajo konkatenativno oblikoslovje,[1] besedne oblike določajo menjave obrazil, najpogosteje pripon ter včasih predpon. V to družino spada večina evropskih jezikov. Primer iz češčine: pridevnik *sladký*

[1]Besede so sestavljene iz več združenih (concatenated) morfemov, ki so predstavljeni v razdelku 2.1.4

Tabela 4.2: Vse besedne oblike za slovensko lemo mesto.

besedna oblika	število	sklon
mest-o	ednina	imenovalnik
mest-a	ednina	rodilnik
mest-u	ednina	dajalnik
mest-o	ednina	tožilnik
mest-u	ednina	mestnik
mest-om	ednina	orodnik
mest-a	množina	imenovalnik
mest-	množina	rodilnik
mest-om	množina	dajalnik
mest-a	množina	tožilnik
mest-ih	množina	mestnik
mest-i	množina	orodnik
mest-i	dvojina	imenovalnik
mest-	dvojina	rodilnik
mest-oma	dvojina	dajalnik
mest-i	dvojina	tožilnik
mest-ih	dvojina	mestnik
mest-oma	dvojina	orodnik

(sladek) lahko spremenimo v *nej-slad-ší-ho* (najslajši – moški ali srednji spol imenovalnik ali tožilnik) z dodajanjem pripone *nej-*, ki predstavlja presežnik, in z menjavo pripone *-ký* (komparativ) s pripono *-ší* ter z dodajanjem pripone *-ho* moški ali srednji spol imenovalnik ali tožilnik.

V tabeli 4.2 je predstavljen primer iz slovenščine za lemo *mesto*, ki vsebuje 18 besednih oblik, za tri števila in 6 sklonov.

4.2 Dvojezični slovar

definicija 4.2.1. *Leksikalne kategorije Lk predstavljajo poljubne podnize obliko-skladenjskih oznak MSD (definirane v definiciji 4.1.2).*

definicija 4.2.2. *Par $\langle lema, oznaka\rangle; oznaka \in Lk; lema \in Lm$; kjer je Lk definirana v definiciji 4.2.1 in Lm definirana v definiciji 4.1.4; označuje lemo s pripadajočim nizom leksikalnih kategorij.*

definicija 4.2.3. *Smer veljavnosti* $smer \in \{LR, RL\}$; $LR \equiv$ *od izvora k cilju;* $LR \equiv$ *od cilja k izvoru; določa kateri prevajalni smeri je namenjen zapis. Izpuščeni operator smeri dopušča poljubno smer.*

definicija 4.2.4. *Prevajalni par* $pp \equiv smer << lI, oI >, < lC, oC >> \leftrightarrow < lC, oC >; lI, lC \in Le; oI, oC \in Lk.$ *Lema izvornega jezika s pripadajočo oznako se prevaja v lemo ciljnega jezika s pripadajočo oznako.*

definicija 4.2.5. *Dvojezični slovar* $Ds : x \in Ds \Rightarrow x \in pp$ *je množica parov iz definicije 4.2.4.*

Definicija 4.2.5 opisuje zapise dvojezičnega slovarja. Oznaka izvorne leme se ponavadi ujema z oznako ciljne leme, še posebej pri sorodnih jezikih. Uporaba oznak omogoča razdvoumljanje lem z istim imenom in različnim pomenom. Dvojezični slovar z menjavo oznak omogoča opisovanje leksikalnih razlik med jezikoma.

```
<e><p>
  <l>okno<s n="samostalnik"/><s n="srednji"/></l>
  <r>prozor<s n="samostalnik"/><s n="moški"/></r>
</p></e>
<e><p>
  <l>okolica<s n="samostalnik"/><s n="ženski"/></l>
  <r>okolina<s n="samostalnik"/><s n="ženski"/></r>
</p></e>
<e><p>
  <l>okoli<s n="predlog"/></l>
  <r>oko<s n="predlog"/></r>
</p></e>
<e><p>
  <l>okoliščina<s n="samostalnik"/><s n="ženski"/></l>
  <r>prilika<s n="samostalnik"/><s n="ženski"/></r>
</p></e>
```

Slika 4.3: Primeri dvojezičnih prevodov lem iz slovenščine v srbščino. Značke so obširneje predstavljene v tabeli 4.1.

Dvojezični slovarji temeljijo na parih *izvorna lema – ciljna lema* oziroma na poravnanih besednih zvezah v lematizirani obliki, torej na dobesednih prevodih lem. Primeri dvojezičnih prevodov lem iz slovenščine v srbščino so predstavljeni na sliki 4.3.

Poleg samega prenosa iz izvornega v ciljni jezik lahko opišemo še prenos oblikoskladenjskih oznak, ki se spremenijo pri samem prevodu. Primer 4.1 kaže prevod slovenske besede *okno* v srbsko besedo *prozor*, pri čemer se spremeni tudi spol iz srednjega v moški.

(4.1) *okno* *prozor*
samostalnik,srednji *samostalnik, moški*
"okno"
"prozor"

4.3 Metode

Vsak modul s slike 3.5 je sestavljen iz osnovne programske opreme in jezikovnoodvisnih podatkov. Podatki so v ogrodju Apertium strukturirani v formatu XML (Bray et al., 2008). Naslednji razdelki predstavljajo opis metod za samodejno izdelavo jezikovnih podatkov iz razdelka 3.4.3.

4.3.1 Izdelava enojezičnih oblikoskladenjskih slovarjev izvornega in ciljnega jezika

Iz oblikoskladenjsko označenega in lematiziranega korpusa najprej izluščimo vse besedne oblike ter jih razdelimo po lemah. Uporabili smo korpus „1984", ki je natančneje predstavljen v razdelku 2.4.1. Primer označene povedi iz tega korpusa je na sliki 2.4. Leme družimo v paradigme, kar nam omogoča sestavljanje manjkajočih besednih oblik; izdelava paradigem je razložena v razdelku 4.3.1.1.

```
lema: cerkev
krn: cerk
primeri besednih oblik:
  besedna oblika: cerkev
  pripona: ev
  MSD: samostalnik ženski spol ednina imenovalnik
  besedna oblika: cerkvah
  pripona: vah
  MSD: samostalnik ženski spol ednina+mestnik
```

Slika 4.4: Del paradigme *cerk-ev*. Lema: cerkev, krn: cerk, dve besedni obliki *cerkev* in *cerkvah*.

Jezikovna gradiva projekta MULTEXT-EAST vsebujejo tudi oblikoskladenjsko označene leksikone vseh podprtih jezikov, med njimi tudi slovenščine in v novi različici tudi srbščine. Metoda omogoča gradnjo takšnih leksikonov, preizkušena pa je bila na jezikovnem paru, ki takšen leksikon sicer vsebuje, vendar jo lahko uporabljamo tudi za ostale jezikovne pare, ki takšnih leksikonov nimajo. Ko je bila izdelana metoda, tudi srbščina še ni imela leksikona. Predstavljena je bila v članku (Vičič, 2009).

4.3.1.1 Izdelava paradigem

Algoritem 1 Algoritem za gradnjo paradigem.

input: $paradigme$
for each $p1 \in paradigme$ **do**
  for each $p2 \in paradigme$ **do**
    if $p1.besednaVrsta = p2.besednaVrsta \ \wedge$
      $\nexists z1 \in p1.zapisi \ \wedge$
      $z2 \in p2.zapisi \ \wedge$
      $z1.MSD = z2.MSD \ \wedge$
      $z1.koncnica \neq z2.koncnica$
    then
      združi paradigmi $p1$ in $p2$ v paradigmo $p1$
      $paradigme \leftarrow paradigme \setminus p2$ ▷ brišemo p2 iz množice paradigem
    end if
  end for
end for
output: $paradigme$

Leme z enakimi spremembami družimo v paradigme; vsaka ima naslednje elemente:

- tipična lema – iz te leme izpeljemo začetno paradigmo;
- krn – najdaljši skupni del vseh besednih oblik v lemi;
- množica vseh besednih oblik, razdeljenih na krn, ter obrazila – k vsaki besedni obliki je zapisana oblikoskladenjska oznaka (MSD) po (Erjavec, 2010).

Primer paradigme je prikazan na sliki 4.4.

Označene leksikone, zbirke besednih oblik s pripisanimi lemami in MSD izvlečemo iz označenega korpusa, kot je „1984”. Paradigme izdelamo z algoritmom 1.

Vse besedne oblike za vsako lemo združimo v razred, ki predstavlja to lemo. Za vsak razred izdelamo paradigmo, ki vsebuje na začetku le zapise ene leme. Sledi združevanje paradigem po algoritmu 1: dve paradigmi združimo v eno, če pripadata isti besedni vrsti (prva kategorija MSD) in če se noben par zapisov ne izključuje. Dva zapisa se izključujeta, če imata MSD in različna obrazila, kot kaže primer na sliki 4.5. Torej dve paradigmi družimo, če ena predstavlja popolno podmnožico druge paradigme. Vsaka paradigma ima shranjen celoten seznam vseh lem, ki jo sestavljajo; ta seznam pri združevanju vsebuje leme obeh paradigem.

Oblikoskladenjski slovarji izvornega in ciljnega jezika so bili zgrajeni s pomočjo paradigem; leme z manjkajočimi besednimi oblikami v originalnih leksikonih so bile dopolnjene, velikost končnega leksikona je bila približno dvajset krat večja od začetnega (Vičič, 2009).

```
lema: cerkev
krn: cerk
besedna oblika: cerkev
pripona: ev
MSD: samostalnik ženski spol ednina imenovalnik
lema: ana
krn: an
besedna oblika: ana
pripona: a
MSD: samostalnik ženski spol ednina imenovalnik
ev != a
```

Slika 4.5: Besedni obliki se ne ujemata, kar pomeni, da paradigmi ne združimo.

4.3.2 Izdelava dvojezičnih prevajalnih slovarjev

Dvojezični prevajalni slovar lahko izdelamo iz poravnanega dvojezičnega korpusa s pomočjo stohastičnih metod oziroma modelov (Vičič, 2008). Poseben problem pri uporabi stohastičnih modelov je v redkih podatkih (sparse data problem) (Katz, 1987). Osnovni korpus ima določeno število dovolj dobro opisanih pravil in dovolj pogosto zastopanih besed, vsebuje pa tudi velik odstotek slabo predstavljenih besed in pravil. Z večanjem korpusa uvajamo tudi nove besede. Tako se odstotek slabo opisanih besed in pravil z večanjem korpusa ne manjša. Problem redkih (pomanjkljivih) podatkov rešujemo s pomočjo naprednih algoritmov, ki upoštevajo predhodno znanje o problemu, izkušnje s sorodnih področij ali pa celo povsem tujih področij. Šumne podatke izločamo s pomočjo zakonitosti v podatkih, z izločanjem ekstremov. Paziti moramo, da pri izločanju napačnih podatkov ne pretiravamo in korpusa preveč ne "porežemo", poenostavimo.

Opisanega problema smo se lotili z dvema na novo razvitima metodama (Vičič in Homola, 2010), ki sta predstavljeni v naslednjih razdelkih:

- *poravnava lematiziranih besed*: iskanje poravnav med lemami jezikovnega para učnega korpusa namesto iskanje povezav med vsemi besednimi vrstami;

- *razširitev dvojezičnega slovarja s podobnicami in iskanje najprimernejših paradigem v ciljnem enojezičnem slovarju*: pri tej metodi se zanašamo na podobnice, kot so predstavljene v razdelku 2.3.5. Leme so prenesene v ciljni jezik brez prevoda, v ciljnem slovarju pa je novi lemi poiskana najprimernejša paradigma.

4.3.2.1 Poravnava lematiziranih besed

Dvojezični prevajalni slovar, opisan je v definiciji 4.2.5, je sestavljen iz parov *izvorna lema z oznako besedne vrste – ciljna lema z oznako besedne vrste*, ki omogočajo prevajanje v ciljni jezik. Oznake besednih vrst v dvojezičnih slovarjih omogočajo enostavno izogibanje dvoumnostim enako imenovanih lem različnih besednih vrst.

```
priti_SAMOSTALNIK biti_GLAGOLP do_PREDLOG
podrt_PRIDEVNIK drevo_SAMOSTALNIK .

o_PREDLOG kateri_ZAIMEK on_ZAIMEK biti_GLAGOLP
praviti_GLAGOL .

...
```

Slika 4.6: Pripravljeni učni podatki: leme in besedne vrste za vsako besedo v korpusu.

Besede v enojezičnih slovarjih so zapisane v lematizirani obliki, besedne oblike pa so zabeležene v paradigmah, kar je natančneje razloženo v razdelku 4.1. Metoda poravnave lem omogoča boljše rezultate v primerjavi s poravnavo besed v korpusu zaradi zmanjšanja prostora iskanja. Omejitev prostora iskanja poveča natančnost modela poravnave besed, vendar v njem ni več informacije o besednih oblikah. To informacijo smo ohranili prek povezave s paradigmami v enojezičnih slovarjih.

Za samo učenje poravnave lem lahko uporabimo poljuben statistični algoritem za iskanje poravnave besed v dvojezičnih, povedno poravnanih korpusih (SMT word-to-word model).

Uporabili smo orodje GIZA++ (Och in Ney, 2003), ki temelji na algoritmu, prikazanem v (Brown et al., 1993). Model je bil naučen na vzporednem, povedno poravnanem seznamu lem z oznakami besednih vrst, ki je bil izluščen iz korpusa „1984”. Del seznama pripravljenih učnih podatkov je prikazan na sliki 4.6. V korpusu je vsaki besedi dopisana še njena MSD in lema. Vzporedno povedno poravnan seznam je bil izdelan tako, da so bile za vsako besedo iz korpusa izluščene samo leme in besedne vrste, ki so prva kategorija MSD.

Oglejmo si še empirični dokaz, da je število besednih oblik v besedilu veliko večje od števila lem, še posebej za visoko pregibne jezike, kot so slovanski. Tabela 4.3 kaže razliko v številu besednih oblik za isti korpus „1984” v petih jezikih, in sicer v treh visoko pregibnih slovanskih jezikih, slovenščini, srbščini in češčini, ter v angleškem in estonskem jeziku, ki sta uporabljena kot referenci.

Tabela 4.3: Primerjava števila lem s številom besednih oblik v korpusu MULTEXT-EAST (Erjavec, 2010). Stolpec razmerje kaže količnik med številom besednih oblik ter lemami.

jezik	število besednih oblik	leme	razmerje
slovenščine	20,923	7,895	2.65
srbščina	21,505	8,392	2.56
češčina	22,273	9,060	2.46
angleščina	11,078	7,020	1.58
estonščina	18,853	8,679	2.17

S predpostavko, da je korpus „1984” dober vzorec opazovanega jezika, v našem primeru slovenščine, dobimo vrednosti števila besednih oblik in lem, kot so predstavljene v tabeli 4.3. Preiskovalni prostor se je za slovenščino v primeru tega korpusa zmanjšal iz 20,923 besednih oblik na 7,895 lem.

4.3.2.2 Razširitev dvojezičnega slovarja s podobnicami in iskanje najprimernejših paradigem v ciljnem enojezičnem slovarju

Ta metoda razširja število vpisov v dvojezični slovar in ustrezno popravi enojezični oblikoskladenjski slovar. Metode, opisane v razdelkih 4.3.2 in 4.3.1, ne zagotavljajo popolne pokritosti enojezičnih slovarjev z dvojezičnim slovarjem. Leme izvornega enojezičnega slovarja, ki po izvajanju teh metod nimajo prevodov, poskušamo prevesti s pomočjo metode, ki temelji na podobnicah. Podobnice so obširneje predstavljene v razdelku 2.3.5. in predstavljajo besede izvornega in ciljnega jezika, ki so si podobne po obliki in pomenu.

Algoritem 2 doda manjkajoče zapise v dvojezični slovar: za vsak vnos izvornega slovarja, ki nima pokritja v dvojezičnem slovarju, tj. nima ustreznega prevoda,

Algoritem 2 Dodajanje manjkajočih zapisov v dvojezični slovar in posledično popravljanje enojezičnih slovarjev.

input: $izvorniSlovar, ciljniSlovar, dvojezicniSlovar$
for each $k \in izvorniSlovar$ **do**
  if $k \notin dvojezicniSlovar$ **then**
    $pIzvorna \leftarrow izvorniSlovar.paradigma(k)$
    $par \leftarrow < k, k >$ ▷ sestavimo par, ki je sestavljen na obeh straneh iz izvorne leme
    $pCiljna \leftarrow ciljniSlovar.minimumDistance(pIzvorna)$ ▷ najdi ustrezno paradigmo v ciljnem slovarju
    $dvojezicniSlovar \leftarrow dvojezicniSlovar \cup \{par\}$ ▷ vstavimo par v dvojezični slovar
    $ciljniSlovar \leftarrow ciljniSlovar \cup \{< pCiljna, k >\}$ ▷ vstavimo izvorno lemo z najverjetnejšo lemo v ciljni slovar
  end if
end for
output: $izvorniSlovar, ciljniSlovar, dvojezicniSlovar$

vstavimo v dvojezični slovar nov par *izvorna lema – izvorna lema*, kar pomeni, da prevajamo lemo v enako lemo v ciljnem jeziku. Poleg same leme je novemu zapisu dodan tudi del MSD; pri primeru na sliki 4.7 je zapisana besedna vrsta in spol. Nov vnos se v ciljni enojezični slovar doda, če nove leme z enako besedno vrsto ne najdemo v ciljnem slovarju. V ciljnem slovarju je dodana nova lema in zanjo izbrana paradigma, ki najbolj ustreza novo dodani lemi; to je paradigma, ki omogoča generiranje besednih oblik z ustreznimi MSD in vsebuje najdaljšo pripono, ki ustreza novi lemi. Algoritem ponovimo še za ciljni slovar, trenutni izvorni slovar postane v drugem delu metode ciljni.

Slika 4.7 s tremi primeri kaže potek dodajanja novih vnosov v dvojezični slovar in v ciljni slovar. Prvi primer kaže slovensko lemo *list* z označeno paradigmo *žvenket/__samostalnik*, ki je prisotna v izvornem slovarju in nima prevoda v dvojezičnem slovarju. Drugi primer kaže nov vnos v dvojezični slovensko-srbski slovar; vpis prevaja slovensko lemo *list* v srbsko lemo *list*. Poleg same leme je zapisan še del MSD, v tem primeru še besedna vrsta (samostalnik) in spol (moški), ki bi se lahko tudi zamenjal, vendar se pri pomanjkanju dodatnih informacij zanašamo na podobnost besed. Tretji primer opisuje nov vpis v ciljnem slovarju. Dodana je nova lema *list* in poiskana najustreznejša paradigma *um /__samostalnik*.

```
slovenski slovar:
lema: list
krn list
paradigma ženket/__samostalnik

slovensko-srbski dvojezični slovar:
list samostalnik moški spol se prevaja v
list samostalnik moški spol

srbski slovar:
lema: list
krn list
paradigma um/__samostalnik
```

Slika 4.7: Razširitev dvojezičnega slovarja. Prvi primer kaže slovensko lemo *list*, ki je prisotna v izvornem slovarju in nima prevoda v dvojezičnem slovarju. Naslednja dva primera kažeta novo dodan vnos v dvojezični slovar in nov vnos v ciljni slovar.

4.3.3 Izdelava statističnega jezikovnega modela ciljnega jezika

Statistični jezikovni model uporablja za prevode v zadnjem delu prevajalnega cevovoda modul za rangiranje kandidatov. Arhitektura omogoča uporabo poljubnega modela, ki ocenjuje kakovost povedi. Zaradi enostavne implementacije algoritmov za izdelavo modela in njegove enostavne uporabe smo se odločili za uporabo statističnega modela jezika, temelječega na trigramih (trigramski model). Teoretične osnove modela so natančneje predstavljene v razdelku 2.1.8, sama implementacija pa v razdelku 3.4.2. Za izdelavo takšnih modelov obstajajo tudi prosto dostopna orodja, kot je CMU Statistical Modelling Toolkit (Clarkson in Rosenfeld, 1997). To orodje smo uporabili pri prvih testih, kasneje pa smo implementirali lastno različico, ki omogoča večjo fleksibilnost in izdelavo popolnoma enakih modelov. Za izdelavo modela smo uporabili lastno učno množico, sestavljeno iz korpusa naključno izbranih člankov iz Wikipedije. Korpus je obširneje predstavljen v razdelku 2.4.3.

4.3.4 Modeliranje oblikoskladenjskih oznak izvornega jezika

Oblikoskladenjska analiza s pomočjo oblikoskladenjskega slovarja izvornim besedam pripiše MSD in leme, vendar so rezultati tega procesa dvoumni. Ponavadi to dvoumnost odpravljamo s pomočjo oblikoskladenjskih označevalcev, ki izberejo najprimernejšo MSD glede na okolico besede. Na ta način je predstavljena večina arhitektur prevajalnih sistemov na osnovi pravil. Naša spremenjena arhitektura (Vičič et al., 2009) opušča oblikoskladenjski označevalnik, izbiro najboljših prevodov pa prepušča poznejšim fazam prevajalnega sistema. Število možnih kandidatov za prevode, če upoštevamo vse dvoumnosti, lahko naraste do neobvladljivih števil; ta pojav je natančneje predstavljen v razdelku 3.4.2.

Algoritem 3 Izločitev vseh nemogočih kandidatov za prevode z uporabo pravil lokalnega ujemanja.

$kandidati \leftarrow$ izdelajKandidate(poved) ▷ izdelaj vse kandidate za izvorno poved
for each $k \in kandidati$ **do**
    $sprememba \leftarrow$uporabi pravila na k
    if $sprememba = k$ **then**
        $koncniKandidati \leftarrow k$ ▷ shrani k med končne kandidate
    end if
end for

Veliko parov oziroma trojk MSD je nemogočih, oznake se glede na okolico medsebojno izključujejo. To dejstvo izkoriščamo pri omejevanju števila kandidatov za prevode. Ujemanje oblikoskladenjskih kategorij lahko modeliramo s pravili, ki temeljijo na regularnih izrazih. Pravila so naučena na enojezičnem označenem korpusu; uporabili smo korpus „1984”. Samodejna izdelava teh pravil je predstavljena v razdelku 5.5. Pravila so opisana v razdelku 5.1. Uporabljena je bila ista oblika pravil kot v Apertiumu. Mehanizem za odkrivanje nemogočih kandidatov prevoda je prikazan v algoritmu 3.

Uporabijo se vsa pravila, katerih regularni izraz lahko apliciramo na kandidata prevoda. Če neko pravilo kandidata spremeni, pomeni, da kandidat ni idealno sestavljen in modul ga zavrže. Tako izbrišemo vse kandidate, ki jih pravila spreminjajo, kar pomeni, da so sestavljeni iz delov, pri katerih se pojavljajo nizi MSD, ki se ne ujemajo, v učnem korpusu pa obstaja dovolj veliko število primerov, ko se enaki nizi ujemajo.

4.4 Krnjenje besednih oblik

Krnjenje – predstavljeno je v razdelku 2.1.1.4 – besedi pripiše njen krn, del besede, ki je skupen vsem besednim oblikam leme. Jezikovna gradiva, ki jih potrebuje prevajalni sistem na osnovi pravil plitkega prenosa, kot je opisan v razdelku 3.2, vsebujejo tudi vsa gradiva, ki jih potrebujemo za izdelavo sistema za krnenje besed. Proces krnenja opišemo z naslednjim postopkom:

- s pomočjo oblikoskladenjskega slovarja analiziramo besedo,
- s pomočjo označevalnika MSD izberemo najprimernejšo oblikoskladenjsko obliko ter njeno lemo,
- v oblikoskladenjskem slovarju pogledamo kakšen krn pripada izbrani lemi.

Algoritem 4 Postopek izdelave krnov.

$oznacenaPoved$ ← analiziraj in razdvoumi izvorno poved ▷ uporabi orodja Apertium
for each $beseda \in oznacenaPoved$ **do**
  l ← poišči lemo za $beseda$ v slovarju
  krn ← poišči krn za l v slovarju
end for

Postopek izdelave krnov je predstavljen na sliki 4. Oblikoskladenjski slovar za vsako lemo vsebuje tudi njen krn in pregibno paradigmo. Kombinacija teh dveh informacij omogoča izdelavo vseh besednih oblik leme. Paradigma krnu dodaja končnice in tako tvori besedne oblike, vsaka besedna oblika v paradigmi ima tudi svojo oblikoskladenjsko oznako. Uporaba oblikoskladenjskega slovarja v prevajalnem sistemu je dvojna, na začetku prevajalnega procesa je uporabljen za oblikoskladenjsko analizo izvornega besedila, vsaki izvorni besedi pripiše vse možne oblikoskladenjske oznake (MSD), ki jih uspe sestaviti iz krnov in obrazil na takšen način, da dobi izvorno besedo. Na koncu prevajalnega procesa je uporabljen oblikoskladenjski slovar za generiranje pravilne besedne oblike glede na lemo in MSD.

Po oblikoskladenjski analizi besedila moramo opraviti le še razdvoumljanje, tj. izbrati najverjetnejšo MSD.

Opisan sistem za krnjenje je bil uporabljen pri evalvaciji sistema s pomočjo metrike METEOR, vrednotenje je predstavljeno v razdelku 7.1.1.2.

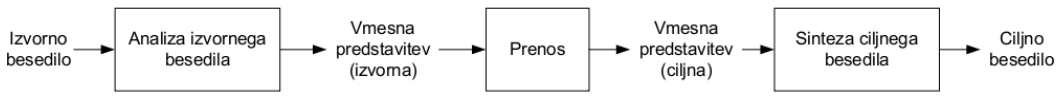
Poglavje 5

Pravila prenosa

V splošnem, povzeto po Hutchins in Somers (1992), lahko delovanje sistemov strojnega prevajanja na osnovi primerov (Rule Based Machine Translation - RBMT) opišemo kot razčlenjevanje (analiziranje) izvornega besedila, ki se zaključi z izdelavo vmesne (simbolne) predstavitve. Iz te predstavitve je v naslednji fazi sintetizirano besedilo v ciljnem jeziku. Prevajalne sisteme opisanega tipa lahko razdelimo glede na vrsto vmesne predstavitve:

- na sisteme z vmesnim jezikom (interlingua) in
- na sisteme s prenosom.

Sistemi z vmesnim jezikom uporabljajo eno vmesno predstavitev, kar omogoča enostavno kombiniranje jezikovnih parov, saj lahko nove prevajalne sisteme izdelamo brez novih pravil za prenos iz izvornega v ciljni jezik, pri čemer za vsak jezik definiramo le pravila za prehod v vmesni jezik. Problem takšnih sistemov pa je, da je izdelava vmesnega jezika zahtevno opravilo še posebej za proste domene in za sisteme z visoko kakovostjo prevajanja, kamor sodijo tudi sistemi za prevajanje med sorodnimi jeziki.



Slika 5.1: Splošna shema sistema za strojno prevajanje s prenosom; sistem ima dve vmesni predstavitvi besedila: izvorno in ciljno, med njima pa poteka prenos.

Sistemi s prenosom uporabljajo dve vmesni predstavitvi, eno za izvorni jezik in eno za ciljni. Takšni predstavitvi je laže izdelati, poleg tega takšen pristop tudi omogoča večjo fleksibilnost, vendar je treba izdelati pravila za prenos za vsak jezikovni par posebej. Primer arhitekture sistemov s prenosom je prikazan na sliki 5.1.

Sistemi strojnega prevajanja s prenosom delujejo tako, da najprej izvedejo oblikoskladenjsko analizo izvornega besedila v vmesno predstavitev izvornega jezika, nato za prenos v vmesno predstavitev ciljnega jezika ter nadalje v dejansko besedilo v ciljnem jeziku uporabijo leksikalne preslikave (dvojezični slovarji) in strukturna pravila prenosa. Nivo analize izvornega besedila in s tem stopnja abstrakcije vmesne predstavitve besedila je odvisen od posameznega sistema za prevajanje in, posredno, od prevajanega jezikovnega para. Prevajanje oddaljenih jezikov, kot sta na primer angleščina in kitajščina, zahteva polno (globoko) analizo (skladenjsko in pomensko). Pri prevajanju sorodnih jezikov, kot je jezikovni par slovenščina-srbščina, pa boljše rezultate dosežemo s plitkim skladenjskim razčlenjevanjem ter posledično s plitkim prenosom in s plitko sintezo (Homola in Kuboň, 2008a; Corbi-Bellot et al., 2005).

5.1 Pravila regularnih izrazov

Pravila na osnovi regularnih izrazov se uporabljajo predvsem v sistemih strojnega prevajanja s plitkim prenosom, shallow-transfer MT. Spremembe s pomočjo teh pravil so najpogosteje povezane z leksikalnimi oblikami; tako je po navadi vmesna predstavitev besedila za vsako besedo sestavljena iz leme, leksikalne kategorije in opisa pregibanj. Takšna pravila so tudi omejena na lokalni kontekst.

5.2 Apertiumov format pravil

Apertiumov modul strukturnega prenosa (Structural transfer module) uporablja za odkrivanje vzorcev fiksne dolžine leksikalnih enot (kosov besedila ali fraz), ki zahtevajo posebno obdelavo glede na slovnične razlike med jezikoma (na primer: spremembe v spolu, sklonu ali številu za zagotovitev ujemanja v ciljnem jeziku, sprememba vrstnega reda besed, leksikalne spremembe, kot na primer spremembe v predlogih ...), tehnologijo končnih avtomatov. Definicija 5.2.1 opisuje zgradbo pravi.

definicija 5.2.1. *Pravila sestavljajo pari* $< vzorec, ukrep >; vzorec \equiv Lk_i \circ b \circ Lk_{i+1} \circ b \circ \ldots; ukrep \equiv akcije(vzorec)$.

Vzorec je predstavljen s sekvenco poljubne dolžine leksikalnih kategorij (definirane v definiciji 4.2.5) izvornega jezika, ločenih s presledki (b – blank). Ukrep določa akcije, ki naj se izvedejo nad sekvencami vzorca ter izhodni vzorec leksikalnih kategorij ciljnega jezika, ki naj se zgradi. Po detekciji vzorcev se izvedejo spremembe, ki so opisane v telesu pravila (izhod modula so spremenjene leksikalne enote). Pravila so zgrajena iz dveh delov: končnega števila elementov, ki opisujejo vzorce fiksne dolžine, in dela, ki omogoča opis akcije, ki je potrebna za spremembo vzorca.

Tabela 5.1: Razlaga značk in atributov zapisa pravil v formatu Apertium.

značka	opis
<rule>	celotno pravilo
<pattern>	vsebuje eno ali več značk (pattern-item), ki definirajo leksikalne oblike, na katere lahko apliciramo pravilo
<pattern-item>	del vzorca, leksikalna enota
<action>	del pravila, ki opisuje ukrep, spremembo vzorca
<let>	sprememba izvornega dela
<clip>	izbere del leksikalne enote, ki ustreza atributom
<lit>	generira niz črk
<lit-tag>	generira niz črk, ki opisujejo jezikovno oznako
<out>	vsebuje vse, kar bo pravilo izpisalo
<lu>	definira vsebino celotne leksikalne enote
<b>	(blank), ločilo med leksikalnima enotama, pogosto je presledek
<call-macro>	klic makra (programske kode)

atribut	opis
side	smer, ki jo naslavlja značka (izvorna/ciljna)
part	ime dela, ki ga naslavlja značka
n	dejanska vsebina značke <pattern-item>
v	dejanska vsebina značk <lit> in <lit-tag>
pos	(position), zaporedna številka leksikalne enote

Vzorci (pattern) so običajno izraženi v leksikalnih kategorijah, na sliki 5.2 *pomožni glagol v prihodnjiku* in *glavni glagol poljubne oblike*. Ukrep (action) določa, katere ukrepe je treba izvesti za ustrezen najdeni vzorec.

Primer pravila je predstavljen na sliki 5.2. Pravilo je sestavljeno iz dveh delov: vzorec *(*pattern) in ukrep *(*action). Opisuje spremembe načina zapisa prihodnjika iz slovenščine v srbščino. Vzorec je sestavljen iz dveh leksikalnih enot: *pomožni glagol biti v prihodnjiku* in *glagol poljubne oblike*, ukrep pa spremeni lemo prvega glagola v *hteti*, obliko prvega glagola v *deležnik* ter obliko drugega glagola v *nedoločnik*; v nadaljevanju so v znački *<lu>* (lexical unit) izpisane leksikalne kategorije

```
<!-- primer: bom kupil-->
<rule>
  <pattern>
    <pattern-item n="pomožni glagol v prihodnjiku"/>
    <pattern-item n="glavni glagol"/>
  </pattern>
  <action>
    <let>
      <clip pos="1" side="tl" part="lema"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="oblika"/>
      <lit-tag v="deležnik"/>
    </let>
    <let>
      <clip pos="2" side="tl" part="oblika"/>
      <lit-tag v="nedoločnik"/>
    </let>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="pomožni glagol"/>
        <clip pos="1" side="tl" part="oblika"/>
        <clip pos="1" side="tl" part="oseba"/>
        <clip pos="1" side="tl" part="število"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="glavni glagol"/>
        <clip pos="2" side="tl" part="oblika"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika 5.2: Primer pravila za strukturni prenos. Pravilo opisuje spremembe načina zapisa prihodnjika iz slovenščine v srbščino. Posamezne značke so predstavljene v tabeli 5.1.

za obe besedi. Posamezne značke zapisa pravil so predstavljene v tabeli 5.1.

Več pravil je predstavljenih v prilogi A.

5.3 Uporaba pravil za strukturni prenos

Modul uporablja za strukturni prenos pri dejanskem prevajanju oblikoskladenjsko označenih leksikalnih enot (po navadi besed ali besednih zvez) pravila prenosa skupaj z dvojezičnim slovarjem. S pravili poskušamo opisati strukturne razlike med jezikoma, torej potrebne spremembe za pravilne prevode iz izvornega v ciljni jezik. Pravila plitkega prenosa, kot jih uporablja Apertium, naslavljajo le dele besedila končne velikosti; večina pravil naslavlja dele besedila dolžine 1, 2 ali 3 besede. Modul v izvornem besedilu poišče dele besedila, ki jih naslavlja pravilo. Pravilo na delu besedila, ki ga naslavlja, izvede akcijo in vrne spremenjeno besedilo.

Sama izbira pokritja posameznih izvornih povedi s pravili poteka po principu najdaljšega ujemanja z leve strani (LRLM – Left-to-Right Longest Match). Za poved v izvornem jeziku je izbrana takšna veriga pravil, da je za dele, pri katerih bi lahko uporabili več pravil, izbrano tisto, ki naslavlja daljše besedilo od leve proti desni.

Primer 5.1 kaže poved „Jutri bom kupil rožo" in njen prevod; del te povedi *bom kupil* je posebej označen in naslavlja pravilo na sliki 5.2.

Oglejmo si še delovanje pravila na primeru 5.1. Prva beseda pokritja, pomožni glagol v prihodnjiku, ustreza besedi *bom* iz primera, druga beseda, glavni glagol, ustreza besedi *kupil*. Pred izvajanjem samega izpisa pravilo postavi novo lemo prvi besedi *hteti* in obliko glagola v deležnik. Obliko drugega glagola spremeni v nedoločnik. Pravilo pri samem izpisu za vsako besedo le izpiše že spremenjene lastnosti v vnaprej dogovorjenem vrstnem redu, kot je prikazano na primeru 5.1.

(5.1)	*bom*	*kupil*
	biti-*glagol p prihod 1os edn*	kupiti-*glagol g delež edn moški*
	"Jutri bom kupil rožo." (SLO)	
	ću	*kupiti*
	hteti-*gl pomožni sedanjik 1os edn*	kupiti-*gl glavni nedoloč*
	"Sutra ću kupiti cvet." (SR)	

5.4 Samodejna izdelava pravil

5.4.1 Izdelava pravil za plitki prenos na osnovi regularnih izrazov

Pri snovanju preizkusa smo se omejili na samodejno izdelavo pravil plitkega prenosa. Takšna pravila so najprimernejša za prevajalne sisteme sorodnih jezikov, kar predstavljajo Homola in Kuboň (2008a) ter Forcada (2006).

Samodejna izdelava pravil plitkega prenosa je bila izvedena s pomočjo orodja, ki je del ogrodja Apertium. Metoda (Sanchez-Martinez in Forcada, 2009) je predstavljena v razdelku 5.4.2.

To orodje izdeluje pravila, ki jih lahko nepredelana uporabimo v ogrodju prevajalnih sistemov, temelječih na Apertiumu (Corbi-Bellot et al., 2005), medtem ko moramo za ostale sisteme format pravil spremeniti, vendar je tudi ta postopek lahko popolnoma samodejen.

5.4.2 Opis metode

Metoda, predstavljena v (Sanchez-Martinez in Forcada, 2009), je osnovana na predlogah poravnave (alignment template – AT), ki so bile predlagane kot izboljšava osnovnih metod statističnega strojnega prevajanja v (Och in Ney, 2004).

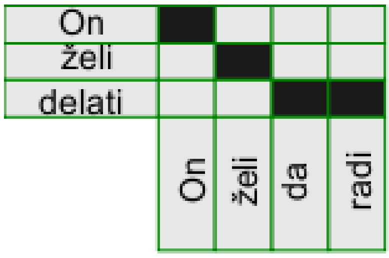
AT lahko definiramo kot posplošitev poravnav parov fraz[1] z uporabo besednih razredov (glsword class), ki so predstavljeni v razdelku 2.1.5.

AT razširimo z množico omejitev, ki nadzorujejo uporabo predlog kot pravil plitkega prenosa. V ta namen:

- povezave med frazami (med deli besedila text chunks), povzete iz učnih primerov, shranimo v dvojezični slovar sistema RBMT, ki omogoča reprodukcijo leksikalne vsebine pri prevajanju;
- besedni razredi so določeni z jezikoslovnim znanjem in ne na podlagi statistike;
- množica omejitev, ki so bile naučene iz učnih primerov, je dodana vsaki AT in omejujejo uporabo AT kot pravila prenosa. Tako spremenjene AT lahko poimenujemo razširjene predloge.

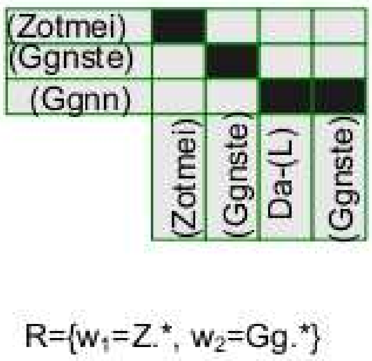
[1]Fraza v okviru tega dela, in tudi splošneje v statističnem strojnem prevajanju, pomeni kolokacijo dveh ali več besed in ne tvori nujno pomensko zaključene enote.

Slika 5.3 kaže poravnano frazo *On želi delati – On želi da radi*; poravnave so v tabeli obarvane črno.



Slika 5.3: Poravnana fraza *On želi delati – On želi da radi*.

Slika 5.4 kaže razširjeno predlogo poravnav (AT), ki je nastala iz poravnane dvojezične fraze s slike 5.3. Besede so zamenjane z besednimi razredi, zapisana je še množica omejitev ciljnega jezika, ki v tem primeru omejuje le prvo in drugo besedo, $*$ predstavlja poljubno zaporedje znakov. Okrajšave oznak MSD so zapisane v tabeli 5.2.



Slika 5.4: Razširjena predloga poravnav. Besede nadomeščajo besedni razredi tako v izvornem kot v ciljnem delu. Dodana je še množica omejitev ciljnega jezika, ki omejuje prvi dve ciljni besedi.

Učenje predlog iz povedno poravnanega dvojezičnega korpusa je sestavljeno iz treh faz:

1. iskanje besednih povezav (word alignments) med izvornim in ciljnim delom

Tabela 5.2: Razširjene kratice, ki so uporabljene na sliki 5.4.

L	členek
Ggnste	glagol glavni nedovršni sedanjik tretja ednina
Ggnn	glagol glavni nedovršni nedoločnik
Zotmei	zaimek osebni tretja moški ednina imenovalnik

korpusa; uporabimo lahko poljubno metodo, najpogosteje pa se uporablja orodje GIZA++, ki temelji na statistični metodi (Och in Ney, 2003);

2. pridobivanje dvojezičnih poravnanih parov fraz; metoda (Och et al., 1999) upošteva vse možne pare fraz do določene dolžine, pri čemer veljajo naslednji pogoji:
   - vse besede si sledijo,
   - nobena beseda dvojezičnega para ni poravnana z besedami izven para,
3. posplošitev teh dvojezičnih parov fraz z uporabo besednih razredov namesto samih besed; posplošitev temelji na zamenjavi posameznih besed z ustreznimi besednimi razredi.

Uporaba razredov besed omogoča opis zamenjave vrstnega reda besed, sprememb pri sklanjanju in uporabi predlogov ter ostalih razlik med izvornim in ciljnim jezikom oziroma njunima vmesnima predstavitvama.

Iz razširjenih predlog lahko sestavimo pravila prenosa.

Strukturni prenos v Apertiumu uporablja končne avtomate za iskanje vzorcev leksikalnih oblik fiksne dolžine. Pri izbiri uporablja algoritem najdaljšega ujemanja vzorca z leve proti desni (LRLM – Left to Right Longest Match). Za izbrana pravila na vzorcu izvede ukrepe (actions), opisane v pravilih. Generično pravilo plitkega prenosa je tako sestavljeno iz vzorca leksikalnih oblik za iskanje in iz opisa potrebnih transformacij na njem. Primer pravila je predstavljen na sliki 5.2.

Pravilo je sestavljeno iz množice razširjenih predlog poravnave (AT) z enakim zaporedjem besednih razredov izvornega jezika, vendar z drugačnimi sekvencami besednih razredov ciljnega jezika in/ali drugačno poravnavo in/ali drugačnim naborom omejitev ciljnega jezika.

Generirana koda izvede AT, ki zadošča omejitvam ciljnega jezika za izbran primer in ki je bila sestavljena iz največjega števila poravnav v učnem korpusu, torej je bila narejena iz največjega števila učnih primerov. Vsako pravilo vsebuje še privzeto AT, ki ima nameščeno najnižjo frekvenco in nima omejitev ciljnega jezika. Ta AT prevaja samo posamezne besede in je uporabljena le, če ni izbrana nobena AT z omejitvami.

5.4.3 Izbira najboljših pravil

Metode za samodejno in nenadzorovano izdelavo pravil po navadi izdelajo veliko množico pravil. Za izbiro najboljših pravil uporabimo metode ocenjevanja pravil. Ena od možnih izbir je naša metoda, predstavljena v (Vičič in Forcada, 2008). Metoda temelji na statističnem trigramskem jezikovnem modelu ciljnega jezika.

Metoda, opisana v nadaljevanju, je bila preizkušena na sistemu za strojno prevajanje sorodnih jezikov jezikovnega para katalonščina-španščina.

Empirično vrednotenje predstavljenih metod je predstavljeno v razdelku 7.2.4.

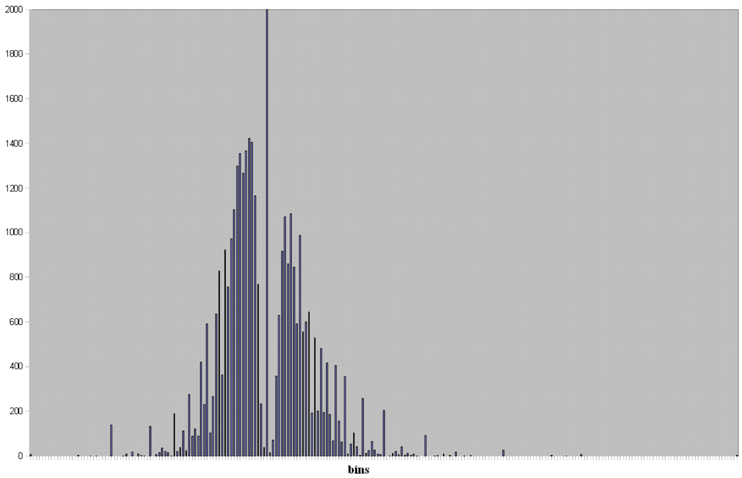
Model ciljnega jezika, kot je opisan v (Bahl et al., 1990; Clarkson in Rosenfeld, 1997), v našem primeru trigramski jezikovni model, se uporablja kot merilo za točkovanje kakovosti izdelanih kandidatov za prevode. Jezikovni model dodeli višjo vrednost, po navadi verjetnost, zaporedjem besed, ki se v učni množici večkrat pojavljajo. Jezikovni model za vsako poved poskuša določiti njeno verjetnost, da bi se pojavila v učnem korpusu, ne pa verjetnost, da je ta poved res prevod izvorne povedi.

Metoda vrednotenja pravil temelji na predpostavki, da model ciljnega jezika zadošča za določitev kakovosti pravil prenosa, saj pravila prenosa v modulu za strukturni prenos prevodom ne spremenijo pomena. Pravila prenosa se največ ukvarjajo z zamenjavo vrstnega reda besed in ujemanjem sosednih besed, kot je opisano v razdelku 5.1.

Kakovost pravil ovrednotimo z dovolj velikim testnim korpusom. Pravila, pri uporabi katerih so prevodi bolje ocenjeni, so tudi sama bolje ocenjena in pri več možnostih sistem izbere pravilo z boljšo oceno.

5.4.3.1 Uporaba jezikovnega modela, ki upošteva dolžino povedi

Trigramski jezikovni model, ki je opisan v prejšnjem razdelku, teži k določanju višje (boljše) vrednosti krajšim povedim, ker obstaja večja verjetnost, da se te pojavijo v učnem korpusu (Bahl et al., 1990; Clarkson in Rosenfeld, 1997). Torej bi v našem primeru model dajal prednost pravilom, ki brišejo besede, tj. pravilom, ki krajšajo povedi. Čeprav ročno napisana pravila le redko odražajo to lastnost, pa tega ne moremo trditi za samodejno grajena pravila. Osnovni metriki smo tako dodali dodaten parameter, ki ta učinek penalizira.



Slika 5.5: Porazdelitev količnika dolžine izvornih povedi v španščini (s – izvor) in njihovih prevodov v katalonščini (t – cilj). Povprečna vrednost je 0,9953 in standardna deviacija je 0,07.

Količnik x dolžine izvornih povedi v španščini in prevodov teh povedi v katalonščini na relativno velikem korpusu El Periódico de Catalunya, ki je predstavljen v razdelku 2.4.6 (okoli 50.000 povedi), je predstavljen na sliki 5.5 in njegovo porazdelitev aproksimiramo z normalno porazdelitvijo:

$$f(x, \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{5.1}$$

kjer je μ povprečje in σ standardna deviacija x. Verjetnost jezikovnega modela pomnožimo s tem penalizacijskim faktorjem.

5.4.3.2 Implementacija metode

Metoda je bila implementirana v ogrodje Apertium, ki je sestavljeno iz modulov, ki delujejo zaporedno, rezultat (izhod) modula je vhod naslednjega modula. Arhitektura je predstavljena v razdelku 3.4. Komunikacija med moduli se izvaja kot preprost vhodno/izhodni tok besedila prek cevi UNIX (UNIX pipes). Umestitev novega modula pomeni samo implementacijo razčlenjevanja izhoda ter izdelavo primernega vhoda za naslednji modul.

Predstavljena metoda je bila implementirana s spremembo že obstoječega modula, modula za strukturni prenos (structural transfer module). Spremenjeni sistem je primeren le za preizkus delovanja metode in za izbiro najboljših pravil, saj spreminja osnovno arhitekturo, kot je predstavljena na sliki 3.2. Spremenjeni modul izdela vsa možna pokritja izvornega oblikoskladenjsko označenega besedila s pravili prenosa. Takšen sistem bi zahteval podobno arhitekturo, kot je predstavljena na sliki 3.5, vendar sprememba ni potrebna, če sistem uporabimo le za izbiro najboljših pravil, ta pravila pa uporabimo v standardnem sistemu.

Danes je lep dan .
je lep dan .
lep dan .
dan .

Slika 5.6: Vse možne pripone povedi *Danes je lep dan*. Zaradi lažje berljivosti so oznake MSD izpuščene.

Vhod modula se razširi na vse možne podnize, iz vhodne povedi izdelamo vse možne pripone, kot je prikazano na sliki 5.6.

Tabela 5.3: Razlaga oblikoskladenjskih oznak španščine in katalonščine.

značka	opis
<det>	determiner – členek
<pos>	positive – pozitiven
<pl>	plural – množina
<sg>	singular – ednina
<n>	noun – samostalnik
<f>	female – ženski
<m>	male – moški
<vbhaver>	verbo haver – pomožni glagol imeti
<vblex>	verbo lexical – navadni glagol
<pri>	present indicative – sedanjik
<p3>	third person – tretja oseba
<pr>	preposition – členek
<pp>	past participle – pretekli deležnik
<def>	definite – določni
<ind>	indefinite – nedoločni
<num>	numeral – število

Izvirni algoritem za strukturni prenos preskoči vse vhodne žetone, ki se ne uporabijo v nobenem od pravil sistema. Za simulacijo te težave je bilo uvedeno posebno prazno pravilo (dummy rule), ki se uporabi vsakič, ko ni pravega pravila z vzorcem, ki se ujema z vhodnim nizom.

Vsaka izdelana pripona, niz besed z oznakami MSD, se obravnava kot posebna poved. Za vsak položaj se uporabijo vsa možna pravila, ne le pravilo z najdaljšim ujemanjem. Za vsak položaj v povedi so shranjena vsa pravila in njihovi rezultati, tj. prevodi za vse vhodne besede, ki jih ta pravila zajamejo. Tak način omogoča izdelavo vseh možnih pokritij vhodne povedi s pravili prenosa.

```
Sus acciones han subido de un 75% desde el ano pa-
sado .
^El seu <det><pos><m><pl>$
^acci<n><f><pl>$
^haver<vbhaver><pri><p3><pl>$
^pujar<vblex><pp><m><sg>$
^de<pr>$
^un<det><ind><m><sg>$
^75<num>$
^des de<pr>$ ^el<det><def><m><sg>$
^any<n><m><sg>$
^passar<vblex><pp><m><sg>$ ^.<sent>$
_0_1_-1_25_-1_0_-1_1_5_25_42
```

Slika 5.7: Delni podatki, trojica, sestavljena iz izvorne povedi, niza pravil (pokritje povedi) in delnega prevoda, ki ga izdela ta niz pravil na izvorni povedi. Značke so obširjene predstavljene v tabeli 5.3.

Vsak niz pravil, ki zajemajo celotno poved, je predstavljen s trojico: vhodna poved, niz pravil in nedokončani prevod, ki ga izdela ta niz pravil. Primer je prikazan na sliki 5.7.

V zadnji fazi metode so uporabljeni vsi preostali moduli prevajalnega sistema, saj je izhod spremenjenega modula prirejen tako, da ostali moduli nemoteno opravljajo svoje delo. Končni prevodi so ovrednoteni s trigramskim modelom ciljnega jezika. Končni rezultat je v obliki peterice, ki je sestavljena iz naslednjih komponent:

1. Izvorna poved.

2. Niz pravil. Pravila so oštevilčena in ta niz predstavlja pravila, uporabljena v tem prevodu. Število -1 označuje, da je bilo uporabljeno prazno pravilo.

3. Rezultat niza pravil na izvorni povedi (z oznakami MSD).

4. Končni prevod z uporabo pravil iz točke 2.
5. Ovrednotenje prevoda iz slike 5.7 z uporabo jezikovnega modela ciljnega jezika.

Peterica z najboljšo oceno predstavlja najboljši niz pravil za izvorno poved.

Sus acciones han subido de un 75% desde el ano pasado .
_0_1_-1_25_-1_0_-1_1_5_25_42
^El seu <det><pos><m><pl>$
^acci<n><f><pl>$
^haver<vbhaver><pri><p3><pl>$
^pujar<vblex><pp><m><sg>$ ^de<pr>$
^un<det><ind><m><sg>$ ^75<num>$
^des de<pr>$ ^el<det><def><m><sg>$
^any<n><m><sg>$
^passar<vblex><pp><m><sg>$ ^.<sent>$
Els seus accions han pujat d'un 75% des de l'any passat .

2.65465883353068E-28

Slika 5.8: Končni rezultat metode je množica ovrednotenih petoric. Vsaka petorica vsebuje izvorno poved, vse pripone te povedi z oblikoskladenjskimi oznakami, niz imen pravil, ki predstavljajo pokritje izvorne povedi s pravili, ciljni prevod in končno oceno jezikovnega modela za ta prevod. Značke so obširjene predstavljene v tabeli 5.3.

5.5 Izdelava pravil na osnovi regularnih izrazov za izražanje lokalnega ujemanja oblikoskladenjskih kategorij

Metoda, opisana v nadaljevanju razdelka, je lastno delo, predstavljeno v (Vičič in Homola, 2010).

Modul za odpravljanje napak lokalnega ujemanja oblikoskladenjskih kategorij uporablja enak tip pravil kot modul za plitki prenos, opisan v (Sanchez-Martinez in Forcada, 2009); pravila so poenostavljena. Pravila odražajo le ujemanje besed, ki so oddaljene največ dve mesti, kot je na primer ujemanje pridevnikov in samostalnikov v sklonu, spolu in številu. Takšna pravila je veliko laže sestaviti kot pravila za strukturni prenos. Metoda odkriva le lokalno ujemanje v okviru največ treh besed (z uporabo trigramskega jezikovnega modela), vendar bi samo z uporabo drugačnega modela njeno delovanje lahko razširili. Zahteve za uporabo metode so preprostejše kot pri metodi za izdelavo pravil strukturnega prenosa, opisani v (Sanchez-Martinez in Forcada, 2009), saj potrebujemo le enojezični, oblikoskladenjsko označeni korpus.

Algoritem 5 Proces samodejne izdelave pravil lokalnega ujemanja iz označenega korpusa.

$trigrami \leftarrow$ izdelaj bigrame in trigrame iz korpusa;
for each $trigram \in trigrami$ **do**
    for each $msd1 \in trigram.besede$ **do**
        for each $msd2 \in trigram.besede$ **do**
            if $msd1 = msd2$ **then**
                $trenutniRazred \leftarrow pravilo(msd1, msd2)$ ▷ ugotovi, kateri deli oblikoskladenjskih oznak v trigramu se ujemajo
            end if
        end for
    end for
    if $r \in vsiRazredi \wedge trenutniRazred = r$ **then**
        $r.count \leftarrow r.count + 1$ ▷ najden razred r z istimi besednimi vrstami in istim ujemanjem
    else
        $vsiRazredi \leftarrow vsiRazredi \cup \{trenutniRazred\}$
    end if
end for
for each $r \in vsiRazredi$ **do**
    if $r.count \leq prag$ **then**
        $vsiRazredi \leftarrow vsiRazredi \setminus \{r\}$ ▷ izbriši razrede s številom kandidatov, ki je manjše od praga
    end if
end for

Pravila lokalnega ujemanja uporabljata dva modula s slike 3.5, in sicer modul za izbiro množice kandidatov in modul za iskanje lokalnega ujemanja. Prvi modul uporablja pravila, naučena na oblikoskladenjsko označenem korpusu izvornih besedil, drugi na enako označenem korpusu ciljnih besedil. Tudi pri tej metodi smo uporabili korpus „1984”, ki vsebuje tako izvorno kot ciljno besedilo.

Tri- in bigrame oblikoskladenjskih oznak, in sicer brez dejanskih besed, torej izključno z oznakami, zgradimo iz označenega korpusa s pomočjo standardne metode za izdelavo stohastičnih jezikovnih modelov, kot je predstavljena v razdelku 2.1.8. Iz osnovnega korpusa „1984” smo najprej izluščili oblikoskladenjske oznake ter na tako pripravljenem korpusu zgradili tri- in bigrame. Tri- in bigrami so bili razdeljeni v skupine glede na besedne vrste (prve dele oblikoskladenjskih oznak) ter na dolžine oznak. V isto skupino so bili dodeljeni tri- in bigrami, ki so vsebovali enake besedne vrste in enako dolge oznake. Ostali deli oznak so bili prosti.

Za vsako leksikalno enoto tri- in bigramov je bilo preverjeno ujemanje oblikoskladenjskih kategorij z vsemi ostalimi enotami. Kandidat za novo pravilo lokalnega ujemanja je shranjen v ekvivalenčni razred z lastnostmi, kot so prikazane na sliki 5.9, če med enotami tri- oziroma bigrama obstajajo ujemanja oblikoskladenjskih kategorij .

```
vzorec naslavljanja za pravilo:
pridevnik, dolžina MSD = 5
samostalnik, dolžina MSD = 4
ujemanja oblikoskladenjskih kategorij: 1-1, 2-2, 3-3
število kandidatov v razredu: 1335
```

Slika 5.9: Ekvivalenčni razred za sestavo pravil lokalnega ujemanja med pridevnikom in samostalnikom. Pridevnik in samostalnik se ujemata v treh lastnostih, ki so na zaporednih mestih 1 = spol, 2 = število, 3 = sklon.

Razred določimo za primernega za izdelavo pravila, če je število kandidatov, ki ga sestavljajo, dovolj veliko in če je število kandidatov v tem razredu, ki ne ustrezajo ujemanjem, zanemarljivo majhno (ocenimo ga kot šum). Prag za izbiro veljavnih pravil je bil določen empirično na podlagi manjšega števila testnih primerov. Metodo za boljšo določitev praga bo treba dodatno raziskati.

Algoritem 5 opisuje ta postopek.

Iz razreda, predstavljenega na sliki 5.9, je bilo izdelano pravilo na sliki 5.10.

5.5.1 Primeri uporabe pravil za lokalno ujemanje oblikoskladenjskih kategorij

Leksikalna podobnost sorodnih jezikov, obširneje predstavljena v razdelku 2.3.4, zagotavlja, da se večina besed pomensko enostavno prevaja v besede v ciljnem jeziku po pravilu "ena-na-ena", torej ena beseda izvornega jezika se prevede v eno besedo v ciljnem jeziku. Tako je izdelava slovarjev enostavna in napake so redke.

```
<rule>
 <pattern>
  <pattern-item n="pridevnik dolžine 5"/>
  <pattern-item n="samostalnik dolžine 4"/>
 </pattern>
 <action>
  <out>
   <lu>
    <clip pos="1" side="tl" part="lema"/>
    <clip pos="1" side="tl" part="pridevnik dolžine5 MSD_0"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_1"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_2"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_3"/>
    <clip pos="1" side="tl" part="pridevnik dolžine5 MSD_4"/>
   </lu>
   <b pos="1"/>
   <lu>
    <clip pos="2" side="tl" part="lema"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_0"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_1"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_2"/>
    <clip pos="2" side="tl" part="samostalnik dolžine4 MSD_3"/>
   </lu>
  </out>
 </action>
</rule>
```

Slika 5.10: Pravilo, zgrajeno iz primerov razreda, prikazanega na sliki 5.9. Pridevnik (prva beseda) se ujema s samostalnikom (druga beseda) v treh delih MSD-oznake, in sicer v kategorijah na mestih 1, 2, 3, ki jih pridevnik povzame po samostalniku. Posamezne značke so predstavljene v tabeli 5.1.

Kljub temu pa obstajajo izjeme, ki jih poskušamo opisati s pomočjo razširitve dvojezičnih slovarjev z oblikoskladenjskimi podatki in s pomočjo pravil.

Oglejmo si dva primera na jezikovnem paru slovenščina-srbščina. Med tema dvema jezikoma obstaja nekaj samostalnikov, ki so različnih spolov v obeh jezikih. Primer je opisan v razdelku 2.3.4 in predstavljen s primerom 2.7. Vsebina primera je zaradi lažje predstavitve ponovno prikazana s primerom 5.2. Pri prevodu besede *okno* v srbsko besedo *prozor* kaže spremembo spola iz srednjega v moški. V obeh jezikih se pridevnik ujema s samostalnikom v spolu, številu in sklonu.

Na sliki 5.11 je zapisano pravilo za lokalno ujemanje pridevnika in samostal-

(5.2)	*Odprto*	*okno.*
	odprto-*pridevnik, srednji spol*	okno-*samostalnik, srednji spol*
	”Odprto okno.” (SLO)	
	Otvoren	*prozor.*
	otvoren-*pridevnik, moški spol*	prozor-*samostalnik, moški spol*
	”Otvoren prozor.” (SR)	

(5.3)	*ptič*	*je*	*letel*
	ptič-*sam, m, ed*	biti-*p. glag, ed*	letete-*glag, m, ed, preteklik*
	”ptič je letel” (SLO)		
	ptica	*je*	*letela*
	ptica-*sam, ž, ed*	jesam-*p. glag, ed*	voziti-*glag, ž, ed, preteklik*
	”ptica je letela” (SR)		

nika, ki popravi napako, ki jo povzroči sprememba spola pri samostalniku.

Pridevnik in samostalnik v pravilu na sliki 5.3 se morata ujemati v spolu, številu in sklonu. Ujemanje je vezano na samostalnik. Ob spremembi teh lastnosti pri prevodu se popravijo oznake pri pridevniku.

Primer 5.3 kaže ujemanje samostalnika, pomožnega glagola leme *jesam – biti* ter glagola v obeh jezikih (SLO, SR).

Na sliki 5.12 je zapisano pravilo za lokalno ujemanje samostalnika, pomožnega glagola leme *jesam – biti* in glagola.

Pravila lokalnega ujemanja so omejena na fiksno določeno okolico in ne zajamejo oddaljenih odvisnosti (long-distance dependencies).

Primer 5.4 kaže poved, pri kateri s pravili lokalnega ujemanja ne moremo popraviti možnih napak. Vrinjeni stavki so poljubne dolžine in jih ne moremo opisati s pomočjo vzorcev končne dolžine, kakršne uporablja Apertium. Sistemi plitkega prenosa takšnih odvisnosti ne zajamejo in pri njihovi uporabi se zanašamo na podobnost med jeziki, pri katerih naj bi bilo takšnih problemov malo.

(5.4) *Kolo,* *bilo* *je* *rdeče,*
kolo-*sam, sr, ed* biti-*p. glag, ed, pret* biti-*p. glag, ed* rdeče-*prid, sr, ed, m*
je *vozilo...*
biti-*p. glag, ed* voziti-*glag, sr, ed, pret*
"Kolo, bilo je rdeče, je vozilo..." (SLO)

Bicikl, *bio* *je*
bicikl-*sam, m, ed* biti-*p. glag, ed, pret* jesam-*p. glag, ed*
crven, *je* *vozilo...*
crven-*prid, m, ed, m* biti-*p. glag, ed* voziti-*glag, sr, ed, pret*
"Bicikl, bio je crven, je vozilo(NAPAKA)..." (SR)

```
<rule>
  <pattern>
    <pattern-item n="pridevnik"/>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="pridevnik"/>
        <clip pos="2" side="tl" part="spol"/>
        <clip pos="2" side="tl" part="število"/>
        <clip pos="2" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="stopnja"/>
        <clip pos="2" side="tl" part="določnost"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="samostalnik"/>
        <clip pos="2" side="tl" part="spol"/>
        <clip pos="2" side="tl" part="število"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika 5.11: Pravilo ujemanja pridevnika in samostalnika, ki si sledita. Besedi se morata ujemati v spolu, sklonu in številu. Pri prevajanju se spreminjajo oblikoskladenjske kategorije samostalnika in ne pridevnika, zato je ujemanje vezano na samostalnik.

```
<rule>
  <pattern>
    <pattern-item n="samostalnik"/>
    <pattern-item n="pomožni glagol jesam (biti)"/>
    <pattern-item n="glavni glagol"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="samostalnik_0"/>
        <clip pos="1" side="tl" part="samostalnik_1"/>
        <clip pos="1" side="tl" part="samostalnik_2"/>
        <clip pos="1" side="tl" part="samostalnik_3"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="pomožni glagol_0"/>
        <clip pos="2" side="tl" part="pomožni glagol_1"/>
        <clip pos="2" side="tl" part="pomožni glagol_2"/>
        <clip pos="1" side="tl" part="samostalnik_2/>
        <clip pos="2" side="tl" part="pomožni glagol_4"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lema"/>
        <clip pos="3" side="tl" part="glavni glagol_0"/>
        <clip pos="3" side="tl" part="glavni glagol_1"/>
        <clip pos="1" side="tl" part="samostalnik_2"/>
        <clip pos="1" side="tl" part="samostalnik_3"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika 5.12: Pravilo ujemanja samostalnika, pomožnega glagola leme *jesam – biti* in glagola. Pomožni glagol in samostalnik se ujemata v številu, samostalnik in glagol na tretjem mestu se ujemata v spolu in številu.

Poglavje 6

Prevajanje na osnovi dreves izpeljave

Statistično strojno prevajanje (SMT), kot je definirano v (Al-Onaizan et al., 1999), predstavlja eno najbolj raziskanih področij CBMT v zadnjih letih. SMT temelji na statističnih modelih, katerih parametri so izpeljani iz opazovanja dvojezičnih vzporednih korpusov. Statistično strojno prevajanje na osnovi dreves izpeljav (statistical machine translation by parsing – SMTbyP), kot je opisano v (Melamed, 2004a), predstavlja podmnožico SMT, v kateri so parametri statističnih modelov naučeni s pomočjo skladenjsko označenih dvojezičnih vzporednih korpusov.

Najpomembnejša prednost sistemov SMTbyP v primerjavi s sistemi SMT je v zmožnosti obvladovanja rekurzivnih struktur v povedih; primer so vrinjeni stavki. Pri učenju modelov in pri samem prevajanju (uporabi modelov) so uporabljeni statistični modeli razčlenjevanja besedila (statistical parsing models). Večina statističnih modelov razčlenjevanja besedila, primera sta (Collins, 2003) in (Charniak, 2000), je naučena na skladenjsko označenih dvojezičnih poravnanih korpusih (treebanks); primer takšnega korpusa je Penn treebank (Marcus et al., 1993). Manj uporabljeni jeziki (less used languages) takšnih korpusov nimajo.

Metoda, predstavljena v nadaljevanju poglavja, omogoča izdelavo samostojnega sistema za strojno prevajanje tipa SMTbyP, medtem ko se ostala poglavja osredotočajo na hitro izdelavo sistemov za strojno prevajanje na osnovi pravil. Najpomembnejši razlog za predstavitev te metode je možnost uporabe takšnih sistemov v primeru nezmožnosti prevodov s pomočjo sistemov na osnovi pravil. Torej nov sistem uporabimo kot pomožni sistem.

6.1 Motivacija

Osnovna hipoteza, na kateri temelji predstavljena lastna metoda (Vičič in Brodnik, 2006), je, da oznake besednih vrst, PoS del oznak MSD, vsebujejo dovolj skladenjskih informacij, ki omogočajo abstrakcijo besed iz učnega korpusa. Statistični model, naučen na besedah korpusa, je modeliran ločeno. Prostor iskanja se z uporabo oznak besednih vrst namesto pravih besed močno zmanjša; tako potrebujemo manj podatkov za učenje učinkovitih prevajalnih modelov. Niz oznak besednih vrst lahko sestavimo iz izvornega označenega besedila, tako da enostavno združimo dele MSD-oznak, ki so pripisane besedam povedi. Takšni nizi predstavljajo liste v drevesih izpeljav, če ne upoštevamo dejanskih besed.

SMTbyP gradi drevo izpeljav izvorne povedi in ga poravna z drevesom izpeljav v ciljnem jeziku. Poravnave so naučene na učni množici, ki je sestavljena iz skladenjsko označenih povedi. Učenje poravnav išče dele drevesnih struktur, ki se pogosto pojavljajo v parih *izvorno skladenjsko označeno drevo – ciljno skladenjsko označeno drevo*.

V svojem pristopu uporabljamo iste algoritme z eno razliko: predstavljena metoda uporablja poravnave med deli nizov oznak besednih vrst izvorne povedi in deli dreves izpeljav v ciljnem jeziku. Tako lahko metodo uporabimo tudi za manj uporabljene jezike. V učnem korpusu, ki vsebuje pare oblike *poved v izvornem jeziku – poved v ciljnem jeziku*, z oznakami MSD označimo izvorne povedi. Iz ciljnih povedi pa zgradimo skladenjsko označena drevesa. Poravnave gradimo s statističnim pristopom, dele nizov oznak besednih vrst poravnamo z deli skladenjsko označenih dreves, če se dovolj pogosto pojavljajo v učnih primerih. Slika 6.3 kaže primer takšne poravnave.

6.2 Predstavitev metode

Večina metod statističnega strojnega prevajanja (Brown et al., 1993; Melamed, 2004a) je jezikovno neodvisna, prevajalne metode delujejo dvosmerno, omogočajo prevajanje iz izvornega v ciljni jezik in obratno. Jezikovna neodvisnost je omogočena z indukcijo prevajalnega znanja iz vzporednih podatkov brez dodatnega jezikovnega znanja.

Osnovna učna množica naše metode (Vičič in Brodnik, 2006) je prav tako poravnan vzporedni dvojezični korpus. Metoda obe predstavljeni splošnosti oziroma neodvisnosti zanemarja, saj zahteva jezik s standardizirano drevesnico v ciljnem jeziku, tj. zbirko skladenjsko označenih povedi (treebank), in jezik s solidnim označevalnikom MSD za izvorni jezik. V prvo skupino sodi le peščica največjih svetovnih jezikov oziroma jezikov, ki so dobro podprti z jezikovnimi tehnologijami, v drugo skupino pa veliko večja množica jezikov, saj je razvoj označevalnika MSD precej manjši problem. Metoda je razdeljena na dva dela, in sicer

- učenje povezav med deli nizov oznak besednih vrst izvornih povedi in deli izdelanih dreves izpeljav ciljnega jezika;

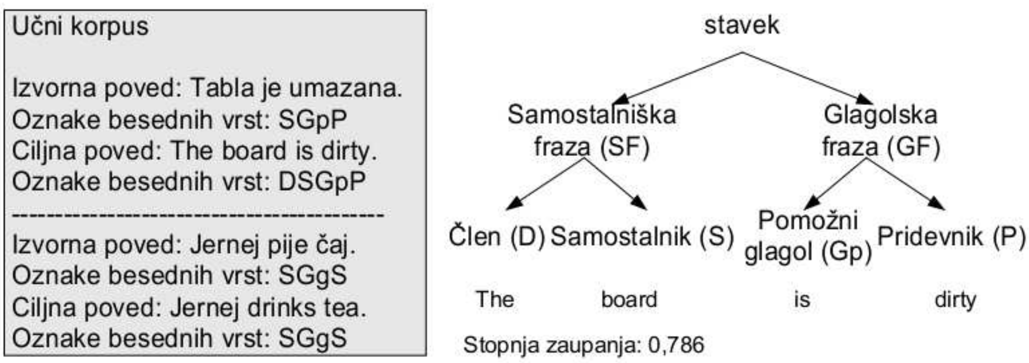
- preiskovanje naučenih primerov iz prvega dela in izdelava množice fiksne velikosti najboljših kandidatov ciljnih prevodov (n-best-set).

6.2.1 Učenje povezav med oznakami in drevesi

Prevajalni model je naučen na dvojezičnem vzporednem korpusu. Korpus je sestavljen iz povedno poravnanih delov izvornega in ciljnega jezika, kot je prikazano na sliki 6.1.

Standardni algoritem SMTbyP gradi drevo izpeljav iz izvorne povedi in ga poravna (poišče skupne točke) z ustreznim drevesom izpeljav v ciljem jeziku. Besede so modelirane v posebnem statističnem modelu, pri čemer lahko uporabimo praktično poljuben model poravnave posameznih besed (word-by-word alignment model).

Naš pristop je v primerjavi z osnovnim različen le v akcijah, ki vključujejo izvorno poved, saj metoda izhaja iz dejstva, da za izvorni jezik ne obstaja dovolj dober algoritem za skladenjsko razčlenjevanje povedi. Vsak par povedi iz učnega dela korpusa je obravnavan samostojno. Ciljna poved je razčlenjena s Collinsovim razčlenjevalnikom (Collins, 2003), ki je bil naučen na skladenjsko označenem korpusu „The Penn Treebank" (Marcus et al., 1993); rezultat razčlenjevanja je drevo izpeljav z izračunano stopnjo zaupanja (confidence score). Primer drevesa izpeljav je predstavljen na sliki 6.1.



Slika 6.1: Učni podatki: oblikoskladenjsko označen korpus in drevo izpeljav, izdelano na podlagi iste povedi. Drevo izpeljav je izdelano le za angleški del poravnanega para, pri slovenskem delu uporabimo le oznake besedilne vrste.

Drevesa izpeljav so sestavljena iz besed v listih, sledijo oznake besednih vrst na prvem nivoju. Oznake besednih vrst so združene v frazah, ki tvorijo preostale nivoje do vrhnjega. Izpuščanje besed v drevesih izpeljave na količino informacij s stališča skladnje skoraj ne vpliva. Notranja vozlišča predstavljajo simbole slovnice.

Skladenjski razčlenjevalnik za izvorni jezik ne obstaja; izvorna poved je oblikoskladenjsko označena, v našem testnem primeru smo uporabili že vnaprej pripravljen in označen korpus „1984”, oznake besednih vrst so bile pridobljene iz korpusa (prvi deli oznak MSD). Opisano zaporedje izdela pare oblike, razvidne iz primera a) na sliki 6.2.

a) <izvorna poved, ciljna poved, izvorne oznake besednih vrst, ciljno drevo izpeljav, stopnja zaupanja za ciljno drevo izpeljav>

b) <izvorna poved, ciljna poved, izvorne oznake besednih vrst, ciljno drevo izpeljav, poravnava, stopnja zaupanja za ciljno drevo izpeljav, točkovanje poravnave>

c) <Tabla je umazana, the board is dirty, SGP, DSGP, (S (SF (D)(S))(GF (Gp)(P))), binarni podatki, 0.786, 0.354>

Slika 6.2: Delni podatki po stopnjah učenja: a) začetni učni podatki; b) končni podatki s poravnavami, ki so točkovane; c) primer s slike 6.1, predstavljen kot končni podatki s primera b), povezave so v binarni obliki.

V naslednji fazi so povezane oznake besednih vrst za vsako poved izvornega jezika z notranjimi vozlišči drevesa izpeljav. Algoritem 6.2.1 poišče pokritje ujemanj med izvornim nizom besednih vrst in najnižjim nivojem drevesa izpeljav, primer poravnave je prikazan na sliki 6.3.

Poravnave so točkovane glede na izbrana pravila, ki so uporabljena pri izdelavi poravnav (vsaka skupina pravil ima svojo težo). Končni izdelek je par, prikazan na sliki 6.3.

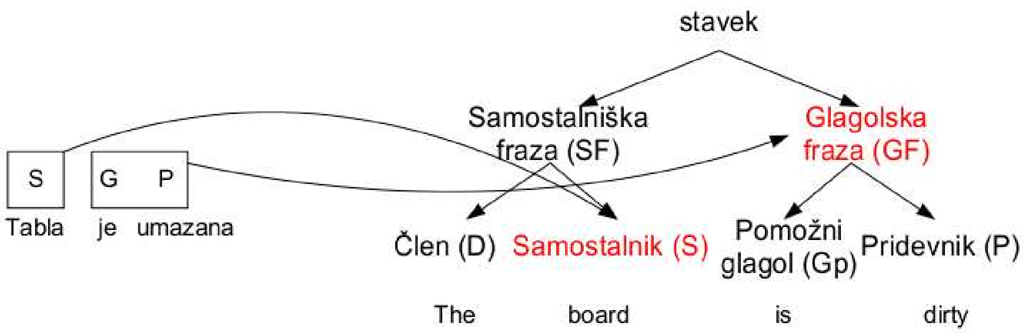
6.2.2 Prevajanje

V tej fazi se odvija prevod vhodne izvorne povedi v poved v ciljnem jeziku. Vhodna poved, poved, ki jo prevajamo, je oblikoskladenjsko označena. Iz teh oznak izberemo le besedne vrste in sestavimo iskalni niz. Algoritem poišče niz oznak besednih vrst v izvornem delu učnih podatkov (za vsako poved). Rezultati so ocenjeni v skladu z uporabljeno iskalno metodo:

- *točno ujemanje*; popolno ujemanje izvornega niza s ciljnim nizom. Ta metoda je ocenjena brez penalizacije.

Algoritem 6 Algoritem poišče pokritje ujemanj med izvornim nizom besednih vrst in najnižjim nivojem drevesa izpeljav, v katerem so prav tako zapisane besedne vrste. Po drevesu se sprehajamo proti korenu, dokler je še zadoščeno kriterijem poravnave ter dokler vozlišče drevesa ne naslavlja celotnega podniza besednih vrst ciljnega jezika, ki je poravnan s trenutnim izvornim podnizom. Postopek ponavljamo do celotnega pokritja izvornega niza besednih vrst.

for each *par in povedi* **do** ▷ vsak par povedi <izvorna,ciljna>
  $izvorneOznake \leftarrow par.izvorna.MSD$;
  $ciljnoDrevoIzpeljav \leftarrow par.ciljna.drevo$;
  while $izvorneOznake.length \geq 0$ **do**
    $a \leftarrow najdaljipodnizizvorneOznake$
    poišči ujemanje med a in najnižjim nivojem $ciljnoDrevoIzpeljav$
    plezaj po $ciljnoDrevoIzpeljav$ do vozlišča v, ki še vključuje celotno ujemanje
    $ujemanja \leftarrow a + v$
  end while
end for



Slika 6.3: Primer poravnave oznak besednih vrst z drevesom izpeljav.

- *ujemanje podobnih nizov*; izvorni in ciljni niz se lahko razlikujeta le za vnaprej določeno uteženo Levenshteinovo razdaljo (Levenshtein, 1965). Penalizacija je odvisna od utežene Levenshteinove razdalje.

- *ujemanje podnizov*; ujemanje izvornega niza z več nizi ciljnega korpusa oziroma ujemanje izvornega dela s podnizom ciljnega niza. Penalizacija je odvisna od števila podnizov in je bila določena le na osnovi majhnega empiričnega testiranja.

Metode so točkovane; osnovno točkovanje je bilo empirično določeno, uporabnik ga lahko spreminja s parametri. Rezultat iskanja je množica najboljših n-teric.

V zadnjem koraku metode se za izdelavo kandidata za prevod samostojno uporablja vsak zapis. Besede ciljnega drevesa izpeljave so napolnjene prek poravnav z izvornimi oznakami besednih vrst in posredno z izvornimi besedami. Izvorne besede so prevedene s pomočjo modela za neposredni prevod besed (word-by-word model).

Prevodi so točkovani s pomočjo točkovanj posameznih faz prevajalnega procesa in pomnoženi z verjetnostjo, da kandidat za prevod sodi v ciljni jezik po statističnem jezikovnem modelu ciljnega jezika; v našem primeru smo uporabili jezikovni model (Clarkson in Rosenfeld, 1997). Kot končni prevod je izbran najbolje točkovan kandidat.

Poglavje 7

Metodologije vrednotenja sistemov in rezultati vrednotenj

Poglavje predstavlja pregled področja vrednotenja sistemov za strojno prevajanje. Omejuje se na prikaz testiranja kakovosti prevodov. Hitrost prevajanja in odzivnost celotnega sistema, prijaznost do uporabnika in ostale lastnosti strojnih prevajalnikov niso predstavljene, saj so opisani sistemi trenutno namenjeni le raziskavam in te vrste funkcionalnost še ni bila upoštevana. Vrednotenje sistema na osnovi dreves izpeljav je opisano v posebnem razdelku, saj rezultati tega vrednotenja niso primerljivi z ostalimi rezultati. Predstavljena je tudi motivacija za izvedbo evalvacije sistemov, omejitve in sredstva, ki smo jih pri tem uporabili, ter sami rezultati.

7.1 Vrednotenje sistemov za strojno prevajanje

Evalvacija je zelo subjektivna in kompleksna, zato univerzalna metoda ocenjevanja še ni določena. Pri strojnem prevajanju se pojavlja veliko metod za vrednotenje sistemov prevajanja. Tolikšno število metod izvira iz dejstva, da se strokovnjaki ne uskladijo, kaj sploh je dober prevod, kaj šele, kakšna so merila za njegovo oceno. Predlagani primeri kriterijev ocenjevanja so predstavljeni v nadaljevanju.

- Hutchins in Somers (1992) navajata tri kriterije:
  - *informativnost* (fidelity, accuracy) pomeni, v kolikšni meri prevod posreduje enake informacije kot izvirnik,
  - *razumljivost* (intelligibility) pomeni, ali je prevod jasen,

  - *ustreznost jezika* (language style) pomeni, ali je v prevodu uporabljen jezik, primeren vsebini in sporočilu.

- Konzorcij LDC (2005) priporoča dva kriterija z izdelanima lestvicama:
  - *vsebinska ustreznost* (translation adequacy) prevodov,
  - *slovnična pravilnost* (fluency) prevodov v ciljnem jeziku.

- Statistični pristopi; vse samodejne metode izvirajo iz te skupine in omogočajo oprijemljivejše ocene. Za vse metode te skupine je skupno, da primerjajo število napak različnih vrst.

- Ocenjevanje dodatnega dela, ki ga je treba vložiti za izdelavo dovolj dobrih prevodov iz rezultatov sistema za strojno prevajanje (post-editing job). Metoda, ki temelji na tej oceni, je predstavljena v razdelku 7.1.6.1.

Vrednotenje MT-sistemov je postalo pomembno področje razvoja MT (Hutchins in Somers, 1992). Ena od možnih delitev metod za vrednotenje (Vičič, 2009) glede na način uporabe je na:

- samodejne metode,
- ročne metode,
- metode, ki vključujejo posege strokovnjakov (ljudi).

7.1.1 Samodejne metode

Samodejne metode temeljijo na primerjavi števila napak različnih vrst. Metode določajo napake kot razlike med prevodom ocenjevanega sistema za strojno prevajanje in enim ali več referenčnimi prevodi.

7.1.1.1 Metrika BLEU

Bilingual Evaluation Understudy – BLEU (Papineni et al., 2001) je bila prva in je še vedno najbolj razširjena metrika za vrednotenje kakovosti prevodov sistemov strojnega prevajanja. Kakovost prevodov je predstavljena kot natančnost ujemanja prevodov sistema za strojno prevajanje z referenčnimi prevodi poklicnih prevajalcev. Vrednosti so izračunane za posamezne prevedene odseke, po navadi povedi, in povprečene za celoten testni korpus. Berljivost in slovnična pravilnost nista upoštevani.

Osnova metrike je primerjava med n-grami kandidata za prevod in referenčnimi prevodi (lahko jih je več), pri čemer ujemanja niso odvisna od položaja. Višje število ujemanj pomeni boljši prevod oziroma prevod, ki je bližje referenčnemu prevodu. Sama primerjava temelji na preciznosti, ki je v paru s priklicem pogosto uporabljana metrika za ugotavljanje pravilnosti metod za iskanje vzorcev.

Za primerjavo kandidata za prevod z enim ali več referenčnimi prevodi uporablja BLEU spremenjeno različico preciznosti. Preciznost lahko predstavimo kot število pravilno klasificiranih elementov (true positives). Sprememba z osnovno različico naj bi poskrbela za lastnost sistemov strojnega prevajanja, ki pogosto težijo k daljšim prevodom, ki jih osnovna različica dobro oceni, sami prevodi pa so neuporabni.

Na sliki 7.1 je predstavljen primer slabega kandidata za prevod in referenčni prevod, ki mu po osnovni različici preciznosti izračunamo visoko vrednost.

```
Hruška je je sladka.
Hruška je sladka.
```

Slika 7.1: Kandidat za prevod in referenčni prevod.

Osnovno preciznost (precision) opisuje enačba 7.1, izračunane vrednosti veljajo za primer na sliki 7.1.

$$P = \frac{m}{w_t} = \frac{4}{4} = 1 \tag{7.1}$$

m predstavlja število besed kandidata za prevode, ki so v referenčnih prevodih, in w_t število vseh besed v kandidatu za prevod. Izračunana vrednost je 1, kar bi pomenilo popoln prevod, kar seveda ni res. Sprememba metrike BLEU je, da algoritem za vsako besedo v kandidatu za prevod poišče največje število pojavitev v referenčnih prevodih m_{max}. Za primer, opisan na sliki 7.1, velja $m_{max} = 1$, saj se beseda *hruška* pojavi le enkrat. Enačba opisuje spremenjeno metriko in izračunano vrednost za primer na sliki 7.1.

$$P = \frac{m}{w_t} = \frac{1+1+1+0}{4} = \frac{3}{4} \tag{7.2}$$

Metoda je uporabljena za n-grame do predefinirane dolžine, po navadi $n = 4$. Rezultati unigramov približno odražajo ustreznost prevodov (adequacy), koliko izvorne vsebine je preneseno v prevod. Rezultati za daljše unigrame pa opisujejo slovnično pravilnost prevoda (fluency).

Metrika BLEU (Papineni et al., 2001) je najbolj razširjena metrika za vrednotenje sistemov strojnega prevajanja, vendar mnogi avtorji, kot na primer (Callison-

Burch et al., 2006) in (Labaka et al., 2007), soglašajo, da BLEU sistematično zapostavlja sisteme RBMT in ni primeren za visoko pregibne jezike. Metrika naj bi bila uporabljana v ožjem obsegu kot doslej, predvsem za primerjanje sorodnih sistemov in za sledenje postopnih sprememb pri gradnji sistema za strojno prevajanje. Za klasično vrednotenje sistemov za strojno prevajanje pa priporočajo metriko METEOR, ki je opisana v razdelku 7.1.1.2.

7.1.1.2 Metrika METEOR

Metric for Evaluation of Translation with Explicit ORdering – METEOR (Banerjee in Lavie, 2005; Lavie in Denkowski, 2009) predstavlja odgovor na pomanjkljivosti trenutno najbolj razširjene metrike za vrednotenje sistemov za strojno prevajanje, BLEU. Metrika temelji na harmonični sredini preciznosti in priklica unigramov (unigram precision and recall), in sicer je priklic močneje utežen kot preciznost. Vsebuje še več metod jezikovnih tehnologij, ki niso prisotne pri ostalih samodejnih metrikah strojnega prevajanja, kot so krnjenje in ujemanje sinonimov kot pomoč pri iskanju ujemanja besed. Krnjenje je predvsem primerno za visoko pregibne jezike, saj omejuje vpliv napačne uporabe pregibanja; na primer napačne uporabe sklona pri samostalnikih. Pomembna razlika z bolj razširjeno metriko BLEU je v tem, da METEOR dobro korelira s človeškim vrednotenjem tudi na nivoju povedi, BLEU metrika pa le na nivoju korpusa (daljšega besedila).

Osnovna enota vrednotenja je poved; algoritem poskuša sestaviti povezave (mappings) med besedami vrednotene in referenčne povedi. Za povezave velja omejitev, da se vsak unigram kandidata za prevod veže z nič ali enim unigramom referenčne povedi in obratno. Poravnave nastajajo v več stopnjah, ki jih nadzorujejo jezikovno osveščeni moduli. Modul je le algoritem ujemanja, ki s pomočjo dodatnega jezikovnega znanja išče boljša ujemanja, na primer modul „wn_synonymy" išče poravnave na osnovi WordNeta (Fellbaum, 1998). Modul „Porter stem" uporablja krnjenje (stemming) pri poravnavi, modul „exact" pa išče le točna ujemanja.

Preciznost P izračunamo po osnovni formuli:

$$P = \frac{m}{w_t}, \tag{7.3}$$

kjer je m število unigramov kandidata za prevod, ki se pojavljajo v referenčni povedi, in w_t število unigramov v kandidatu za prevod. Priklic R izračunamo po osnovni formuli:

$$R = \frac{m}{w_r}, \tag{7.4}$$

kjer je m število unigramov kandidata za prevod, ki se pojavljajo v referenčni povedi, in w_r število unigramov v referenčni povedi. Harmonska sredina preciznosti in priklica je izračunana z devet krat večjo utežjo priklica po:

$$F_{mean} = \frac{10 * P * R}{R + 9 * P}. \tag{7.5}$$

Opisano harmonsko sredino preciznosti in priklica penaliziramo z vrednostjo p, ki opisuje kongruenco daljših nizov med kandidatom za prevod in referenčno povedjo. Unigrame grupiramo v najmanjše število nizov, za katere velja, da so si unigrami sosedni v kandidatu za prevod in referenčni povedi. Čim daljša so zaporedja medsebojno povezanih unigramov kandidata za prevod ter referenčne povedi, manjše število nizov dobimo; prevod, ki je enak referenčnemu prevodu, bo imel le en niz. Vrednost p izračunamo po enačbi:

$$p = 0.5 * (\frac{c}{u_m})^3, \tag{7.6}$$

kjer je c število nizov in u_m število unigramov, ki so bili uspešno povezani. Končna ocena je izračunana kot M v enačbi 7.7

$$M = F_{mean} * (1 - p). \tag{7.7}$$

7.1.1.3 Metrika WER

Stopnja napačnih besed (word error rate – WER) je bila ena prvih statističnih metod za določanje kakovosti prevodov. Temelji na uteženi Levenshteinovi razdalji (weighted Levenshtein edit-distance) (Fu, 1982). Ta predstavlja razširitev osnovne razdalje (Levenshtein, 1965), ki šteje najmanjše število sprememb, ki jih moramo opraviti med prevodom sistema za strojno prevajanje in referenčnim prevodom. Število sprememb še utežimo z dolžino povedi. Dovoljene spremembe so vstavitev, brisanje in zamenjava besede.

Izračun vrednosti WER za eno poved je oblike:

$$WER = \frac{S + D + I}{N}, \tag{7.8}$$

kjer je

- *S* – število substitucij,
- *D* – število izbrisov,
- *I* – število vstavkov,
- *N* – število besed v povedi.

Predstavljena metrika opisuje velikost napake prevajalnega sistema. Pogosto želimo rezultate takšnega vrednotenja predstaviti kot kakovost prevajalnega sistema; takrat uporabimo različico metrike, ki predstavlja stopnjo prepoznanih besed (word recognition rate – WRR), ki je enostavno razlika med „popolnim" prevodom in napako sistema:

$$WRR = 1 - WER. \tag{7.9}$$

7.1.2 Ročne metode

7.1.3 Vrednotenje po smernicah LDC

Smernice LDC (LDC, 2005) so bile predstavljene na letni delavnici o vrednotenju strojnega prevajanja NIST (NIST machine translation evaluation workshop) in so najpogosteje uporabljena načela za ročno ocenjevanje kakovosti prevodov sistemov za strojno prevajanje. Pri ročnem ocenjevanju kakovosti prevodov upoštevamo dve lestvici, ki predstavljata vsebinsko ustreznost prevodov (adequacy) in slovnično pravilnost prevodov v ciljnem jeziku (fluency).

Prva lestvica kaže kakovost prevodov, tj. koliko izvornega pomena se je pri prevodu ohranilo:

- 5 = vse,
- 4 = večina,
- 3 = precej,
- 2 = malo,
- 1 = nič.

Druga lestvica kaže slovnično pravilnost povedi v ciljnem jeziku. Pri prevodu v ciljni jezik velja:

- 5 = prevod brez napak,
- 4 = dober ciljni jezik,
- 3 = ciljni jezik, kot nematerni jezik (non-native language),
- 2 = ciljni jezik z veliko napakami,
- 1 = nesmiselno besedilo.

Ločeni lestvici za kakovost prevodov in slovnično pravilnost sta bili izdelani ob predpostavki, da lahko tudi prevod z veliko slovničnimi napakami prikaže vso informacijo, ki je zapisana v originalu.

7.1.4 Vrednotenje po smernicah ALPAC

Leta 1966 je Alpac (Automatic Language Processing Advisory Committee) objavil raziskavo (ALPAC, 1966), ki velja za prvi večji poskus vrednotenja strojnega prevajanja. Ocenjevali so prevode iz ruščine v angleščino, in sicer z vidika razumljivosti (intelligibility) in z vidika zvestobe (angl. fidelity). Ocenjevalci so bili pred raziskavo posebej šolani. Raziskava je pokazala, da so bile razlike med ocenjevalci majhne, kljub temu pa priporočajo, da pri vrednotenju sodelujejo vsaj trije ali štirje ocenjevalci.

7.1.5 Vrednotenje po smernicah DARPA

Pri agenciji DARPA (Defense Advanced Research Projects Agency) so objavili metodologijo vrednotenja prevajalnih sistemov (Baker et al., 1992). Glavni izziv pri vrednotenju je bil zmanjšati subjektivnost, ki jo lahko merimo s stopnjo odstopanja med ocenjevalci. Najprimernejše metode, ki so jih izbrali za nadaljnjo uporabo, so vključevale vrednotenje razumljivosti s pomočjo testov razumevanja, vrednotenje primernosti, ki so jo izvedli materni govorci angleščine, in vrednotenje, ki temelji na kriterijih, kot jih predlagata Hutchins in Somers (1992).

7.1.6 Metode, ki vključujejo posege strokovnjakov

7.1.6.1 Utežena Levenshteinova razdalja

Metrika, temelječa na uteženi Levenshteinovi razdalji (weighted Levenshtein edit-distance) (Fu, 1982), poznana tudi kot Word Error Rate (WER), ki je natančneje predstavljena v razdelku 7.1.1.3, izračuna najmanjše število sprememb, ki jih moramo narediti za izdelavo *pravilne* povedi v ciljnem jeziku iz samodejno izdelane povedi (prevoda ocenjevanega sistema).

Kot pravilen prevod po navadi pojmujemo poved, ki popolnoma izraža pomen izvorne povedi in je v ciljnem jeziku zapisana slovnično pravilno. Opisana metrika kaže, koliko dela moramo opraviti za izdelavo dobrega prevoda iz že izdelanega strojnega prevoda. Metrika v grobem ponazarja kompleksnost opravila končnega čiščenja prevodov (post-editing task).

Izvedbo testiranja kakovosti prevodov prevajalnega sistema s pomočjo utežene Levenshteinove razdalje sestavljajo naslednja dejanja:

- izbira testnih povedi v izvornem jeziku;
- prevajanje testnih povedi s pomočjo testiranega sistema;
- *ročno* popravljanje prevodov (popravljavci upoštevajo navodilo čim manjšega števila sprememb);
- izračun utežene Levenshteinove razdalje.

Tako kot pri metriki WER tudi tukaj uporabimo različico, ki predstavlja kakovost prevodov, WRR = (1 - WER).

7.2 Rezultati

Metode, predstavljene v razdelku 4.3, se osredotočajo na gradnjo sistemov za strojno prevajanje za sorodne, oblikoslovno bogate jezike. Poleg same uporabnosti predstavljenih metod in ocenjevanja hitrosti izdelave novih prevajalnih sistemov, stremi predstavljeno vrednotenje k preverjanju kakovosti samodejno izdelanih podatkov za sisteme strojnega prevajanja na popolnoma funkcionalnih sistemih.

Zgrajeni in ovrednoteni so bili štirje popolnoma delujoči sistemi za strojno prevajanje:

1. SL-SR, prevajalni sistem za jezikovni par slovenščina-srbščina;
2. SL-CS, prevajalni sistem za jezikovni par slovenščina-češčina;
3. SL-EN, prevajalni sistem za jezikovni par slovenščina-angleščina;
4. SL-ET, prevajalni sistem za jezikovni par slovenščina-estonščina.

7.2.1 Opis sistemov

Sistem za prevajanje jezikovnega para slovenščina-srbščina (SL-SR) je bil zgrajen kot pilotni sistem, ki je služil za testiranje naših metod v procesu svojega razvoja. Metode, predstavljene v tem prispevku, so bile pregledane v več iteracijah (sistematične napake so bile odpravljene, popravki pa vključeni v novo iteracijo sistema). Ta jezikovni par je bil uporabljen za preverjanje kakovosti predstavljenih metod na popolnoma funkcionalnem sistemu strojnega prevajanja. Oba jezika sta pregibno, oblikoslovno in derivacijsko bogata. Čeprav sta oba jezika sorodna, visoka stopnja pregibnosti obeh jezikov še vedno zahteva oblikoskladenjsko analizo izvornega jezika in posledično oblikoskladenjsko sintezo ciljnega jezika.

Sistem za prevajanje jezikovnega para slovenščina-češčina (SL-CS) je bil izdelan za preverjanje uporabnosti metod, predstavljenih v razdelku 4.3, na novem jezikovnem paru sorodnih jezikov in z namenom, da bi preizkusili, kako hitro je mogoče izdelati nov sistem. Lastnosti tega jezikovnega para so podobne lastnostim prvega jezikovnega para (SL-SR).

Sistema za jezikovna para SL-EN in SL-ET sta bila izdelana za oceno uporabnosti predstavljenih metod in celotne arhitekture za oddaljene jezikovne pare. Rezultati, predstavljeni v razdelkih 7.2.2.1, 7.2.2.2 in 7.2.2.3, kažejo jasno zmanjšanje kakovosti prevodov z uporabo enake metodologije in enakih učnih podatkov. Estonski jezik je bil izbran kot oddaljen pregibni jezik, angleški jezik pa kot oddaljen izolativni jezik, tj. jezik, pri katerem je razmeroma malo pregibanj. Za postavitev vsakega od predstavljenih sistemov smo porabili enako časa, in sicer je sisteme izdelala ena oseba z uporabo običajnega osebnega računalnika[1] in za vsak sistem porabila po dva delovna dneva.

7.2.2 Izbrane metrike vrednotenja

Ovrednotenje prevodov je bilo opravljeno s tremi metodami vrednotenja, vsaka od njih je podrobno opisana v razdelku 7.1, sama uporaba pa v nadaljevanju razdelka:

1. Samodejno objektivno vrednotenje z uporabo metrike METEOR (Banerjee in Lavie, 2005; Lavie in Denkowski, 2009).
2. Vrednotenje z metodo, ki vključuje posege strokovnjakov na podlagi utežene Levenshteinove razdalje.
3. Vrednotenje z metodo, ki vključuje posege strokovnjakov na osnovi predlaganih smernic (LDC, 2005).

Metrike BLEU nismo uporabili, saj po mnenju več avtorjev ni primerna za takšno evalvacijo; razlogi so širše predstavljeni v razdelku 7.1.1.1.

7.2.2.1 Samodejno objektivno vrednotenje z metriko METEOR

Metrika METEOR je natančneje opisana v razdelku 7.1.1.2. Uporabljena je bila javno dostopna implementacija metrike METEOR (Lavie in Denkowski, 2009), različica v0.6. Metrika kot enega izmed algoritmov za večanje korelacije s človeško

[1] prenosni računalnik z 2 GB RAM in procesorjem Intel Core2 duo.)

oceno za visoko upogibne jezike uporablja mehanizem krnjenja. Uporabili smo mehanizem krnjenja, ki je stranski izdelek našega prevajalnega sistema, ki je obširneje predstavljen v razdelku 4.4. Rezultati so predstavljeni na sliki 7.2, števila, označena z *, kažejo vrednosti metrike METEOR z običajnim Porter-stem (Porter, 1980) algoritmom krnjenja, ostala števila kažejo vrednosti metrike z lastnim algoritmom krnjenja.

Kot testna množica primerov s poravnanimi referenčnimi prevodi je bil uporabljen večjezični vzporedni korpus JRC-Acquis (Steinberger et al., 2006). Nov korpus je bil uporabljen zaradi prevelike koreliranosti povedi v učnem korpusu, vse povedi so iz istega romana (Orwell, 1949). Izbira korpusa JRC-Acquis (Steinberger et al., 2006) je bila pragmatična, to je še edini korpus, ki vsebuje vse jezike, ki smo jih uporabili v eksperimentu. Testni primeri so bili še posebej izbrani, odločili smo se za omejitev dolžine povedi na 40 besed, saj je korpus JRC-Acquis specifičen in vsebuje veliko povedi z naštevanjem odsekov pravnih besedil. Manjkajoče besede iz testnih povedi so bile ročno dodane v enojezični in prevajalni slovar. S tem postopkom smo se želeli izogniti napakam tipa „besede izven domene" (out of domain error).

Testna množica primerov je bila sestavljena iz 1500 povedi. Rezultati vrednotenja so predstavljeni na sliki 7.2.

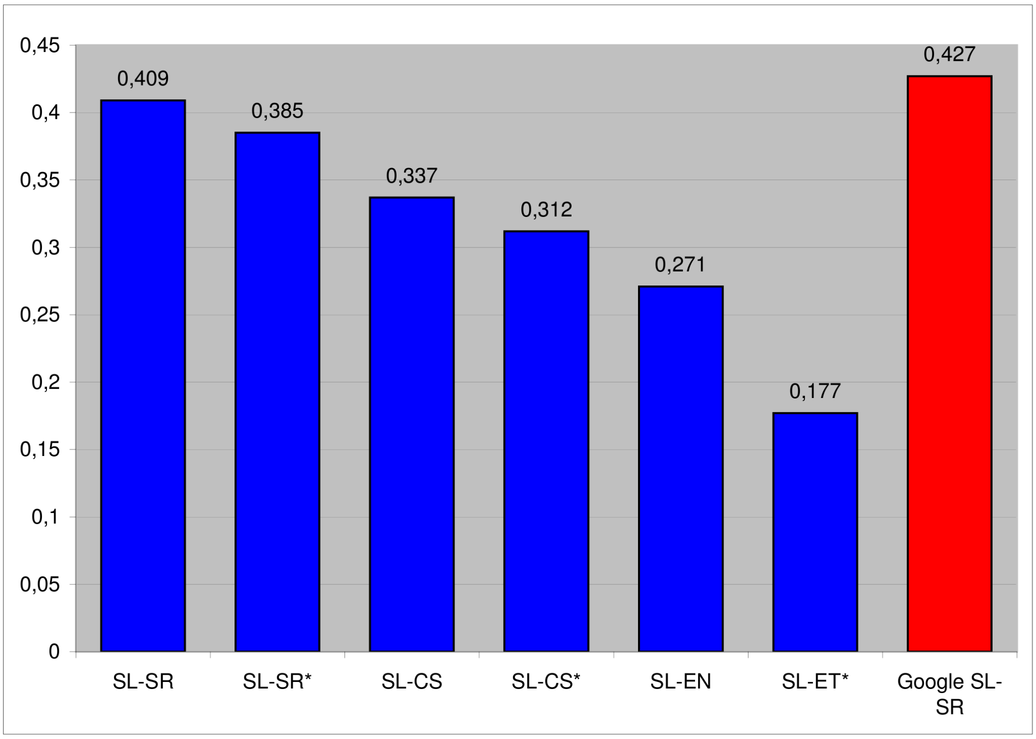
Zadnji stolpec na sliki 7.2 kaže vrednosti evalvacije na istih testnih podatkih za prevajalni sistem Google Translate (Google, 2008) in prevajalni par slovenščina-srbščina. Opis sistema in način vrednotenja je predstavljen v (Vičič, 2010). Vrednosti so primerljive, sistem na osnovi pravil pa lahko še izpopolnimo z ročnim pregledom.

Rezultati Googlovega prevajalnega sistema, na sliki 7.2, so boljši kot rezultati našega sistema. Moramo se zavedati, da je tudi naš sistem popolnoma samodejno grajen in omogoča lingvistom veliko izboljšav, saj lahko vsa gradiva dodatno spreminjamo oziroma dodajamo nova, pri Googlovem sistemu pa je to praktično nemogoče. Edini način izboljšave je uporaba večje količine učnih gradiv oziroma izdelava drugačnih algoritmov.

7.2.2.2 Vrednotenje z metodo, ki vključuje posege strokovnjakov na podlagi utežene Levenshteinove razdalje

Utežena Levenshteinova razdalja je natančneje predstavljena v razdelku 7.1.6.1. Iz korpusa smo naključno izbrali 200 povedi, ki niso bile del učne množice. Uporabili smo soležne povedi za vse jezike (iste testne primere, vendar v drugem jeziku).

Povedi so bile prevedene s prevajalnim sistemom in ročno popravljene. Kot

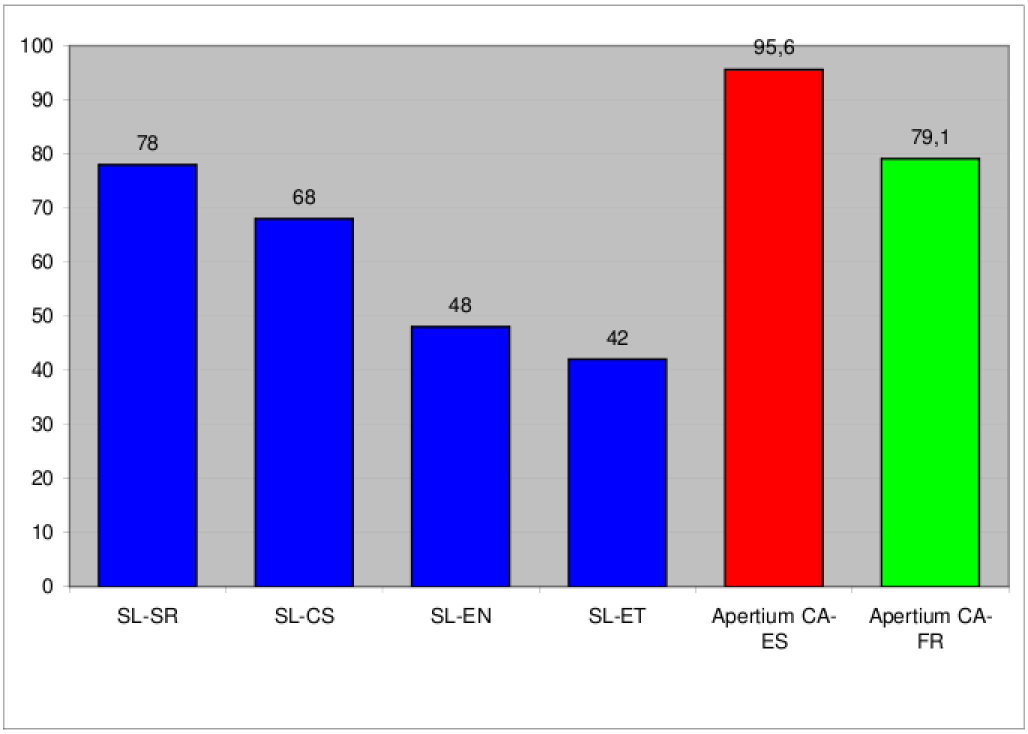


Slika 7.2: Rezultati vrednotenja z metriko METEOR. Uporabili smo korpus Acquis Communautaire (Erjavec et al., 2005). Ovrednotenja, označena z zvezdico *, predstavljajo uporabo krnjenja z algoritmom Porter-stem, ostala pa z uporabo lastnega algoritma za krnjenje.

zadovoljivi prevod štejemo prevod, ki popolnoma odraža vsebino izvorne povedi in je v ciljnem jeziku slovnično popolnoma pravilno zapisan. Med prevodi sistema in popravljenimi prevodi smo izračunali uteženo Levenshteinovo razdaljo. Rezultati so prikazani na sliki 7.3 in predstavljajo WRR, Word Recognition Rate (1 - WER), ki odraža kakovost sistema namesto njegove napake.

Popravljavci so sledili napotkom, naj prevode popravijo s čim manj spremembami. Vrednotenja so večinoma opravljali študenti in raziskovalci, sodelujoči pri poskusu. Ocene kakovosti prevodov slovenščine in češčine sta opravila po dva ocenjevalca, ki jima je bil ciljni jezik materni jezik (native speaker). Ocene kakovosti prevodov angleščine, srbščine in estonščine je opravil po en ocenjevalec, ki mu je bil ciljni jezik materni jezik.

Zadnja dva stolpca na sliki 7.3 kažeta vrednosti ročno izdelanih sistemov za



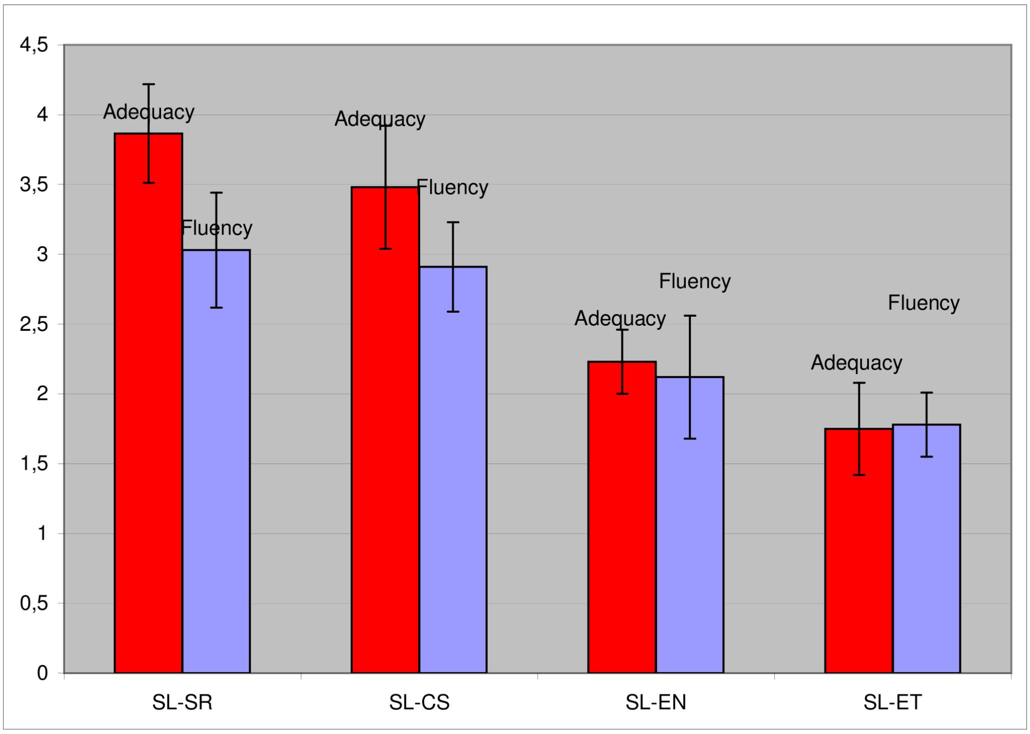
Slika 7.3: Rezultati vrednotenja s pomočjo metrike Word Recognition Rate (WRR).

strojno prevajanje na osnovi Apertiuma za dva različna jezikovna para. Opisi sistemov in rezultati so obširneje predstavljeni v (Villarejo et al., 2010). Najvišje vrednosti sistemov, postavljenih s testirano metodo, so primerljive z nižjimi vrednostmi ročno postavljenih sistemov.

7.2.2.3 Vrednotenje z metodo, ki vključuje posege strokovnjakov na podlagi vnaprej podanih smernic

Metoda ročnega vrednotenja na podlagi smernic (LDC, 2005) je natančneje predstavljena v razdelku 7.1.3. Iz korpusa smo naključno izbrali 100 povedi, ki niso bile del učne možice. Uporabili smo soležne povedi za vse jezike (iste testne primere, vendar v drugem jeziku).

Vrednotenja so večinoma opravljali študenti in raziskovalci, sodelujoči pri poskusu. Ocene za slovenski in češki jezik sta opravila po dva ocenjevalca, ki jima je bil ciljni jezik materni jezik (native speaker), ocene za angleški, srbski in estonski jezik je opravil po en ocenjevalec, ki mu je bil ciljni jezik materni jezik. Rezultati so predstavljeni na sliki 7.4. Rezultati za sistema SL-SR in SL-CS so zadovoljivi, predvsem vrednosti za ustreznost prevodov, vrednosti za preostala sistema so nižje, predpostavljamo, da predvsem na račun različnosti jezikovnih parov.



Slika 7.4: Rezultati vrednotenja po smernicah (LDC, 2005). Povprečne vrednosti dveh neodvisnih ocenjevanj kažejo visoke vrednosti za vsebinsko ustreznost prevodov (adequacy) in nižje vrednosti za slovnično pravilnost.

Tabela 7.1 kaže zadovoljivo (satisfactory) (SL-CS) in zelo visoko (very-high) (SL-SR) ujemanje med ocenjevalci (inter-rater agreement) glede na Cohenov koeficient kapa (Cohen, 1960), ki je predstavljen v nadaljevanju. Opisna magnituda rezultatov je povzeta po (Landis in Koch, 1977).

Cohenov koeficient kapa (Cohen, 1960) je statistično merilo ujemanja med ocenjevalci (inter-rater agreement). Na splošno velja prepričanje, da je ta mera robustnejša od enostavnega deleža ocenjevanj, ki se ujemajo, saj upošteva ujemanja, ki

Tabela 7.1: Cohenov koeficient kapa (Cohen, 1960) za sistema SL-SR in SL-CS kaže zadovoljivo ujemanje (satisfactory agreement) za jezikovni par (SL-CS) ter znatno ujemanje (substantial agreement) za jezikovni par (SL-SR). Pričakovano ujemanje je ujemanje, pri katerem bi se ocenjevalca odločala naključno. Opazovano ujemanje je enako koeficientu kapa. Vsi ocenjevalci so ocenjevali po 100 primerov.

ff	sl-sr	sl-cs
kapa	0,86	0,69
opazovano ujemanje	$0,86$	$0,69$
pričakovano ujemanje	$0,300$	$0,317$
število primerov	100	100

se pojavljajo po naključju. Cohenov koeficient κ meri ujemanje med dvema ocenjevalcema, ki klasificirata po N elementov v C medsebojno izključujočih se razredih:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{7.10}$$

kjer je $Pr(a)$ relativno opazovano ujemanje med ocenjevalcema, $Pr(e)$ je hipotetična verjetnost naključnega ujemanja, ki je določena iz podanih podatkov, če bi vsak ocenjevalec elemente ocenjeval naključno.

Če se ocenjevalca popolnoma ujemata, je κ = 1, če pa se ujemata le v številu primerov, kot bi pričakovali, če bi odgovarjala naključno, je κ = 0. V literaturi so se pojavile smernice, ki določajo magnitudo κ; ena od možnih razdelitev je podana v (Landis in Koch, 1977), ki deli vrednosti κ po naslednjem kriteriju:

- < 0; ni ujemanja (no agreement),
- 0,00-0,20; rahlo ujemanje (slight agreement),
- 0,21-0,40; dokajšnje ujemanje (fair agreement),
- 0,41-0,60; zadovoljivo ujemanje (moderate agreement),
- 0,61-0,80; znatno ujemanje (substantial agreement),
- 0,81-1,00; skoraj popolno soglasje (almost perfect agreement).

7.2.3 Vrednotenje sistema na osnovi dreves izpeljav

V empiričnem vrednotenju sistema za strojno prevajanje na osnovi dreves izpeljav sta bili obravnavani dve težavi:

- kakovost prevodov nizov oznak besednih vrst;
- stopnja uspešnosti iskanja nizov besednih vrst.

Vsaka naloga je podrobneje predstavljena v 6. poglavju.

7.2.3.1 Eksperimentalno okolje

Za postavitev testnega okolja so bila uporabljena že dostopna orodja, kjer je bilo le možno. Kar nekaj aplikacij pa je bilo ustrezno spremenjenih. Izdelan je bil tudi nov modul, ki implementira metodo, predstavljeno v 6. poglavju. Modul je bil umeščen v sistem GenPar. Sledi kratek opis testnega okolja po komponentah:

- Za postavitev prevajalnega sistema je bil kot osnova uporabljen GenPar, obširneje je predstavljen v razdelku 3.3.
- Metrika, ki temelji na uteženi Levenshteinovi razdalji (weighted Levenshtein edit-distance), obširneje je predstavljena v razdelku 7.1.1.3, je bila uporabljena za določanje kakovosti nizov oznak besednih vrst.
- Kot učni in testni korpus je bil uporabljen korpus „1984"; učni in testni podatki se niso prekrivali.
- Za učenje modela poravnave besed je bil uporabljen korpus SVEZ-IJS (Erjavec, 2006); natančneje je predstavljen v razdelku 2.4.4.

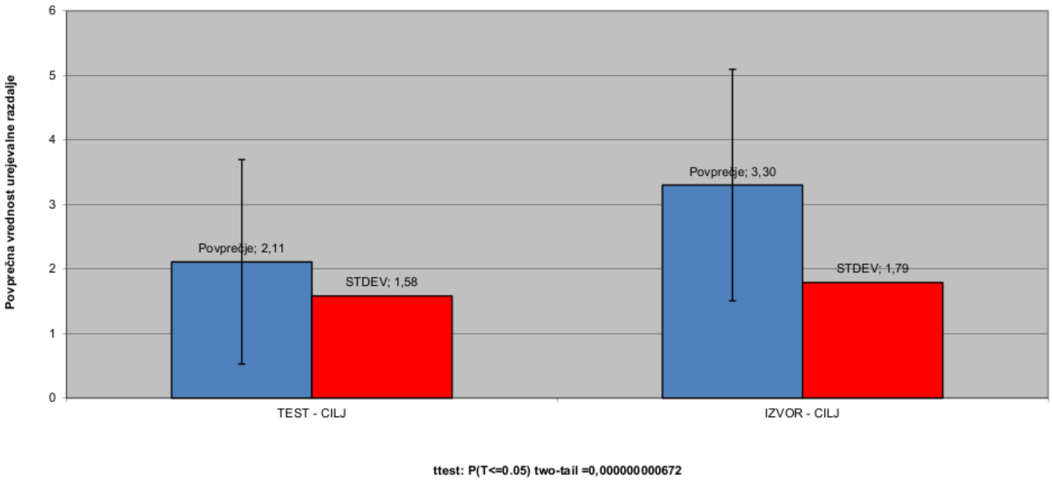
7.2.3.2 Nabor podatkov

Korpus „1984" je že oblikoskladenjsko označen, izluščen je bil le prvi del oznake MSD, tako da ni bilo potrebe po uporabi oblikoskladenjskega označevalca. Zaradi časovne zahtevnosti je bil uporabljen le del celotnega korpusa, in sicer povedi, ki so krajše od 15 besed.

Pred učenjem je bil korpus razdeljen na dva dela, na učno ter testno množico in sicer v razmerju $9 : 1$, torej učna množica je bila $9 - krat$ večja od testne. Kot testni vhodni podatki prevajalnega sistema, so bili uporabljeni nizi oznak besednih vrst izvornega jezika (IZVOR). Kot referenčne vrednosti v postopku ocenjevanja so bili uporabljeni nizi besednih vrst ciljnega jezika (CILJ). Izhod sistema, ciljni niz oznak besednih vrst (TEST) je bil primerjan z nizoma (IZVOR) ter (TEST).

7.2.3.3 Rezultati

Kakovost prevedenih nizov oznak besednih vrst. Vrednotenje kakovosti prevedenih nizov oznak besednih vrst, kandidatov za končne prevode, je bila izvedena z uporabo metrike utežene Levenshteinove razdalje (Fu, 1982). Vsak niz oznak besednih vrst predstavlja list v drevesu izpeljav in je posledično osnova za končni prevod (oznake besednih vrst zamenjamo z dejanskimi besednimi oblikami s pomočjo izvornih besed).



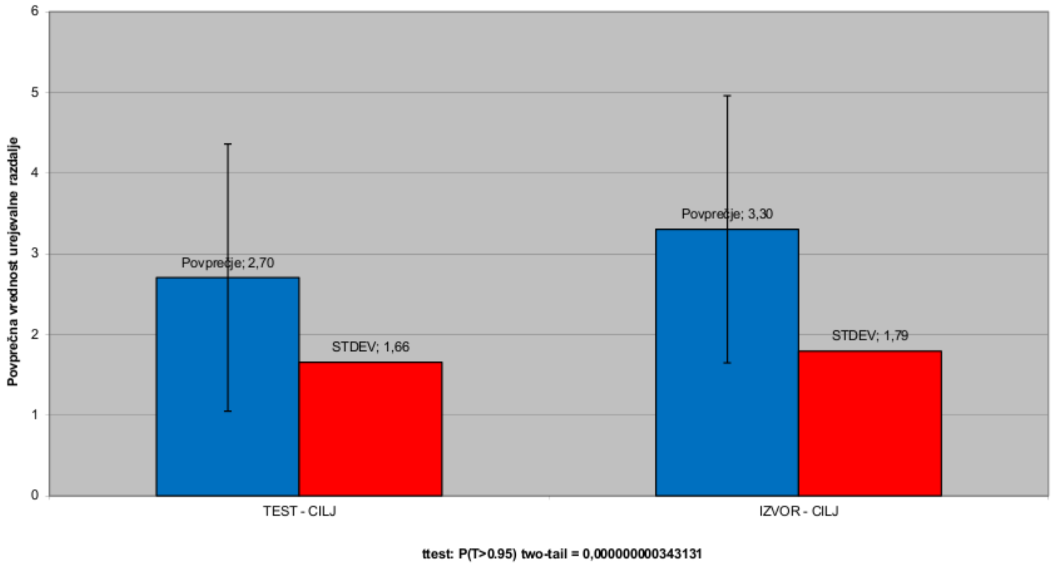
Slika 7.5: Kakovost najdenih nizov oznak besednih vrst z metodo, ki upošteva le nize z urejevalno razdaljo 0. T-test kaže signifikantno razliko med povprečnima vrednostma.

Urejevalna razdalja (edit distance) kaže, koliko se razlikujeta niz, ki ga je izdelal sistem (TEST), in referenčni niz (CILJ). Manjše vrednosti pomenijo boljše rezultate.

Urejevalna razdalja med izvornim nizom besednih vrst (IZVOR) ter referenčnim prevodom (CILJ) kaže, koliko naj bi se spremenila struktura povedi pri prevodih. Vrednosti so bile izračunane kot osnovne vrednosti, ki jih želimo izboljšati (baseline values). Izvedeni sta bili dve skupini testov:

- Kakovost ciljnih nizov besednih vrst, ki jih je vrnil sistem z metodo, ki je upoštevala le urejevalno razdaljo 0 (edit distance = 0). Rezultati vrednotenja so prikazani na sliki 7.5.

- Kakovost ciljnih nizov besednih vrst, ki jih je vrnil sistem z metodo, ki je upoštevala urejevalne razdalje, manjše od 3 (edit distance < 3). Rezultati vrednotenja so prikazani na sliki 7.6.



Slika 7.6: Kakovost najdenih nizov oznak besednih vrst z metodo, ki upošteva le nize z urejevalno razdaljo manjšo kot 3. T-test kaže signifikantno razliko med povprečnima vrednostma.

T-testa na slikah 7.5 in 7.6 kažeta, da sta bili povprečji uteženih urejevalnih razdalj med referenčnimi nizi (CILJ) in nizi, ki jih je izdelal sistem (TEST), za obe metodi signifikantno nižji kot povprečji uteženih urejevalnih razdalj med izvornimi (IZVOR) in referenčnimi nizi. To pomeni, da metoda pri strukturnem prenosu vnaša signifikantno količino informacij.

Stopnja uspešnosti iskanja nizov oznak besednih vrst. Drugi problem, ki je bil obravnavan v empiričnem testiranju, je bil oceniti stopnjo uspešnosti iskanja niza oznak besednih vrst; tj., koliko kandidatov za prevode je algoritem dejansko našel.

Če izvorni niz oznak besednih vrst, ki ga sestavimo s pomočjo označevalnika MSD, v izvorni učni množici ni najden niti ob upoštevanju urejevalne razdalje, se osnovni prevajalni tok konča. Za sam prevod potrebujemo drugo metodo, lahko bi povečali urejevalno razdaljo. Tabela 7.2 kaže delež vhodnih povedi, ki imajo v učni množici vsaj enega kandidata za prevod. Delež strmo naraste pri mejni vrednosti urejevalne razdalje 2 in vse povedi imajo kandidate za prevode pri urejevalni razdalji 5. Vzorčenje je bilo izvedeno trikrat (test1, test2, test3).

Tabela 7.2 kaže, kako se odstotek najdenih nizov besednih vrst veča z višanjem

Tabela 7.2: Delež testnih povedi, ki imajo v učni množici vsaj enega kandidata za prevod.

Urejevalna razdalja	test1	test2	test3
0	40%	37%	39%
1	43%	42%	44%
2	67%	69%	70%
3	89%	91%	89%
4	95%	97%	97%
5	100%	100%	100%

Tabela 7.3: Kakovost nizov oznak besednih vrst glede na izbran prag urejevalne razdalje. Kakovost je ovrednotena s povprečno urejevalno razdaljo do referenčnih prevodov.

Iskalni algoritem	Povprečna razdalja	Standardna deviacija
prag = 0	2,11	1,96
prag = 1	2,34	1,89
prag = 2	2,70	1,66
prag = 3	3,20	1,89
prag = 4	4,05	2,34
prag = 5	5,11	3,01

praga za urejevalno razdaljo. Z višanjem praga urejevalne razdalje se zmanjšuje kakovost nizov oznak besednih vrst in posledično kakovost prevodov. Ta pojav kaže tabela 7.3. Algoritem s pragom urejevalne razdalje, večje od 0, tako uporabimo le v skrajni sili; prag povečujemo v minimalnih korakih in vrednosti, večje od 2, niso priporočljive. Povprečna napaka rezultatov iskalnega algoritma s pragom velikosti 3 je skoraj enako velika kot napaka izvornih nizov, torej metoda s tem algoritmom rešitev ne izboljšuje. Večji prag pa rešitve občutno poslabša.

7.2.4 Vrednotenje metode za izbiro najboljših pravil

Dve vrsti preskusov sta bili izvedeni z uporabo metrik, predstavljenih v razdelku 5.4.3. Opisni rezultati so predstavljeni v tabeli 7.4, kumulativni in primerjalni rezultati pa so predstavljeni v tabeli 7.5.

Poleg pravil iz delujočega prevajalnega sistema so bila izdelana še posebna pravila, ki namerno uvajajo napake v prevode. S pomočjo teh pravil smo želeli poka-

Vzorec: členek presledek samostalnik presledek pridevnik Ukrep: izhod: členek presledek pridevnik

Slika 7.7: Zlonamerno pravilo, ko se pojavi vzorec členek – samostalnik – pridevnik; pravilo izpiše le členek in pridevnik (pobriše samostalnik).

Original:
Sus acciones han subido de un 75 % desde el ano pasado.
Les seves accions han pujat d'un 75 % des de l'any passat.
Changed:
Sus acciones han subido de un 75 % desde el ano pasado.
Els seus accions han pujat d'un 75 % des de passatpassa-tany.

Slika 7.8: Primer vpliva pravila s slike 7.7. Prvi par kaže primer španske povedi in primernega prevoda v katalonščino, drugi par pa prevod, ki ga je pravilo pokvarilo.

zati, ali predstavljene metrike odkrijejo takšna, slaba pravila. Uporaba jezikovnih modelov za odkrivanje teh slabih pravil je eden od osnovnih ciljev te raziskave. Zlonamerno pravilo je uporabljalo vzorec, ki se v besedilih pogosto pojavi; v testnem korpusu se je pojavil v 12 % povedi. Pravilo je predstavljeno na sliki 7.7 in izdela krajše povedi. Primer vpliva tega pravila je prikazan na sliki 7.8.

Tabela 7.4 kaže opisne rezultate ocenjevanja. Testiranje je bilo izvedeno na množicah 480, 1000 in 2000 povedih španskega dela testnega korpusa.

Tabela 7.5 predstavlja rezultate vrednotenja sistema z uporabo nekoliko spremenjenega nabora pravil, ki je vseboval tudi zlonamerna pravila. Namen tega preskusa bi bil ugotoviti, ali je za odkrivanje slabih pravil mogoče uporabiti verjetnostne porazdelitve.

7.2.4.1 Raziskava algoritmov za izbiro pravil

Pri uporabi sistema z nespremenjenimi pravili bi obe metriki zamenjali le majhen odstotek pokritij, ki jih določi algoritem LRLM (nekaj primerov na vsako testno množico). Podrobnejši pregled primerov kaže, da so bila alternativna pokritja izbrana zaradi zamenjave spola, ki ima v korpusu večjo verjetnost (v tem primeru moški spol namesto ženskega); ta sprememba prinaša napačne povedi. Rezultati kažejo, da algoritem LRLM v večini primerov najde optimalno pokritje izvornih

Tabela 7.4: Rezultati vrednotenja. Testiranje je bilo izvedeno na množicah 480, 1000 in 2000 povedih. Stolpec *popravek dolžine (da/ne)* kaže, katera metoda je bila uporabljena: metrika z uporabo osnovnega trigramskega jezikovnega modela (ne) ali metrika z uporabo spremenjenega modela (da). Drugi stolpec kaže, koliko pokritij algoritma LRLM sta metriki označili kot neoptimalna. Tretji stolpec kaže to število v odstotkih.

Popravek dolžine	# ne-optimalnih pokritij LRLM	odstotek napak
480 testnih povedi		
ne	4 (vse napačne)	$0,80\%$
da	4 (vse napačne)	$0,80\%$
1000 testnih povedi		
ne	4 (vse napačne)	$0,40\%$
da	4 (vse napačne)	$0,40\%$
2000 testnih povedi		
ne	12 (vse napačne)	$0,60\%$
da	6 (vse napačne)	$0,40\%$

povedi. Ta rezultat je bilo pričakovati, saj se algoritem LRLM uporablja v sistemih za strojno prevajanje in so pravila napisana zanj.

Tabela 7.5: Rezultati vrednotenja. Testiranje je bilo izvedeno na množicah 480, 1000 in 2000 povedih. Stolpec *popravek dolžine (da/ne)* kaže, katera metoda je bila uporabljena: metrika z uporabo osnovnega trigramskega jezikovnega modela (ne) ali metrika z uporabo spremenjenega modela (da). Drugi stolpec kaže, koliko pokritij algoritma LRLM sta metriki označili kot neoptimalna. Tretji stolpec kaže število napačno označenih pokritij. Četrti stolpec kaže število pokritij algoritma LRLM, ki uporabljajo zlonamerno pravilo, peti stolpec, pa kaže odstotek povedi, ki jih je metoda odkrila in ki uporabljajo zlonamerno pravilo (idealno naj bi bile odkrite vse takšne povedi).

Popravek dolžine	# ne-optimalnih pokritij LRLM	število napak	# ne-optimalnih pokritij LRLM z zlonamernim pravilom	v %
480 testnih povedi (60 povedi vsebuje zlonamerno pravilo)				
ne	35	4	31	52%
da	52	4	48	80%
1000 testnih povedi (120 povedi vsebuje zlonamerno pravilo)				
ne	65	4	61	50%
da	90	4	95	79%
2000 testnih povedi (218 povedi vsebuje zlonamerno pravilo)				
ne	109	11	98	40%
da	178	10	168	77%

Poglavje 8

Razprava in nadaljnje delo

8.1 Zaključki

Delo predstavlja poskus združevanja več metod za hitro postavitev prevajalnih sistemov za sorodne visoko pregibne jezike. Sistem temelji na skupini strojnega prevajanja na osnovi pravil plitkega prenosa, ki se je na pilotnih sistemih, predstavljenih v tem delu in v mnogih znanstvenih člankih, kot so (Corbi-Bellot et al., 2005; Hajič et al., 2003; Homola, 2010; Scannell, 2006), izkazala kot najprimernejša za postavitev sistema za strojno prevajanje sorodnih jezikov. Metode so bile preizkušene na primeru samodejne izdelave prevajalnega sistema. Vrednotenje kaže perspektivne rezultate, čeprav je možnost napredovanja še vedno dovolj velika.

Za preizkus predstavljenih metod so bili izdelani sistemi za strojno prevajanje in izvedeno je bilo vrednotenje kakovosti prevodov teh sistemov. Uporabljene so bile tri metrike, ena popolnoma samodejna ter dve metriki, ki zahtevata ročno vrednotenje oziroma ročno popravljanje prevodov.

Hitrost delovanja posameznih modulov novih sistemov je enaka hitrosti sistemov, ki so bili ročno sestavljeni, saj opisane metode sestavljajo jezikovna gradiva, sami algoritmi prevajanja pa so nespremenjeni. Edina razlika je uvedba nove arhitekture, ki temelji na množici kandidatov za prevode. Tako se časovna kompleksnost celotnega sistema poveča za faktor velikosti množice kandidatov, torej za konstantni faktor n, ki ga uporabnik sam določi. Pri izvajanju večine vrednotenj smo se odločili za $n = 50$, kar pomeni 50-kratno upočasnitev prevajalnega sistema. Čas za postavitev samih sistemov, torej čas za izvajanje opisanih metod, ki se izvedejo le enkrat, je odvisen od velikosti učnih podatkov. Kot je opisano v 7. poglavju, so programi jezikovna gradiva izdelali čez noč na osebnem računalniku. Rezultati vrednotenj kažejo primerljive rezultate z ročno izdelanimi sistemi. Rezultati metrike METEOR so primerljivi z rezultati vrednotenja dveh ročno grajenih sistemov, rezultati metrike WRR pa so primerljivi z rezultati vrednotenja prevajalnika Google.

Zavedati se moramo, da lahko sisteme na osnovi pravil še izpopolnimo z ročnim pregledom prevajalnih gradiv in ročno identifikacijo napak. Ena največjih prednosti uporabljene tehnologije je prav možnost izboljšave postavljenega prevajalnega sistema, saj eksplicitno zapisana pravila prenosa in slovarji omogočajo iterativno izboljševanje kakovosti prevodov. Prikazane metode omogočajo hitrejšo izdelavo sistemov za strojno prevajanje za nove jezikovne pare.

Metoda za izbiro najboljših pravil za strukturni prenos omogoča izbiro najboljših pravil iz množice samodejno grajenih pravil. Metoda ni primerna za izbiro ročno grajenih pravil, saj so ta pisana tudi za posebne primere, ki jih statistične metode pogosto spregledajo. Ročno pisana pravila so pisana z mislijo na uporabljeni izbirni algoritem in tako prirejena načinu izbire. Časovna kompleksnost predstavljene metode je primerljiva s časovno kompleksnostjo prevajalnega sistema Apertium, ki je približno 5000 besed na sekundo. Celotna metoda se je tako izvajala 1500 sekund.

Metoda je bila preizkušena na relativno majhnem korpusu velikosti 1700 povedi. Rezultati potrjujejo uporabnost metode. Kakovost ciljnega niza besednih oznak (z uporabo urejevalne razdalje) kaže na statistično pomembne razlike v osnovnih vrednosti in vrednosti, ki smo jih pridobili z uporabo predstavljene metode. Časovna kompleksnost predstavljene metode je primerljiva s časovno kompleksnostjo originalnega sistema oziroma celo manjša, saj so uporabljene metode enostavnejše.

Iskanje nizov oznak besednih vrst lahko ne najde iskanega niza v učnih podatkih, delež takih primerov je še vedno visok. Z urejevalno razdaljo (zlasti velikosti 2 ali manj) se stopnja uspešnosti sicer poveča, zmanjšuje pa se kakovost najdenih nizov ciljnih oznak vrst.

8.2 Nadaljnje delo

Prag za izbiro veljavnih pravil je bil določen empirično na podlagi manjšega števila testnih primerov. Metodo za boljšo določitev praga bo treba dodatno raziskati.

Potrebno je vrednotenje sistema na korpusu, ki vsebuje daljše povedi, saj je bilo vrednotenje izvedeno na korpusu s krajšimi povedmi (do dolžine 15).

Zadnja stopnja prevajalnega procesa, uporaba modela poravnave posameznih besed (word-by-word alignment model), še vedno ni implementirana. Poleg same implementacije ostaja neizdelana še izboljšava samega algoritma za poravnavo besed pri smtByP, predvsem z uvedbo prevajanja fraz.

Poleg metod za hitrejšo izdelavo jezikovnih gradiv za nove prevajalne sisteme, lahko sisteme za nove jezikovne pare sestavimo iz že obstoječih sistemov, primer je prikazan v (Homola in Vičič, 2010).

Ena od možnosti nadaljevanja predstavljenega dela je postavitev splošnega prevajalnega sistema za južnoslovanske jezike, torej sistema za strojno prevajanje vseh uradnih jezikov bivše Jugoslavije, saj so si ti jeziki med sabo sorodni.

8.3 Prispevki k znanosti

Disertacija predstavlja izvirne prispevke s področja strojnega prevajanja naravnih jezikov na osnovi pravil plitkega prenosa. Izvirne prispevke k znanosti smo objavili v naslednjih glavnih publikacijah (Vičič in Brodnik, 2008; Homola et al., 2009; Mikolič et al., 2009; Vičič, 2009) in v vrsti referatov, objavljenih na mednarodnih konferencah (Vičič in Erjavec, 2002; Vičič in Brodnik, 2006; Vičič, 2007a,b; Vičič in Forcada, 2008; Vičič, 2008; Vičič et al., 2009; Homola in Vičič, 2010; Vičič in Homola, 2010). Razdelek predstavlja zbrani seznam najpomembnejših izvirnih prispevkov k znanosti s krajšimi opisi. Obširneje so prispevki predstavljeni v ločenih razdelkih.

1. Metoda za prevajanje s pomočjo dreves izpeljave za jezike z omejeno podporo jezikovnih tehnologij, jezike, za katere ne obstaja standardna drevesnica (treebank), zbirka skladenjsko označenih povedi, kot je Penn treebank (Marcus et al., 1993).

   Vsebina je širše predstavljena v 6. poglavju. Metoda je bila predstavljena v (Vičič in Brodnik, 2006) in (Vičič in Brodnik, 2008).

2. Metoda za samodejno izdelavo oblikoskladenjskih slovarjev:

   - samodejno označevanje paradigem;
   - samodejno luščenje paradigem za visoko pregibne jezike in izdelava pripadajočih leksikonov;
   - samodejna izdelava dvojezičnih prevajalnih slovarjev.

   Širše je predstavljena v 4. poglavju.

   Prispevek je predstavljen v (Vičič, 2007a), (Vičič, 2007b) in (Vičič, 2009).

3. Ocenjevanje pravil za strukturni prenos:

   - uporaba ocenjevanja pravil,
   - algoritmi za izbiro pravil,
   - metrike ocenjevanja pravil.

Prispevek je predstavljen v (Vičič in Forcada, 2008).

4. Hitra izdelava prevajalnega sistema na osnovi RBMT za sorodne jezike. Kot empirična preizkušnja predstavljenih metod je bil izdelan delujoč prevajalni sistem za jezikovni par slovenščina-srbščina, Guat. V sistemu so implementirani in preizkušeni opisani prispevki k znanosti. Izbrane so metode za samodejno izdelavo jezikovnih gradiv za postavitev prevajalnega sistema. Guat temelji na odprtokodnih tehnologijah, tudi vse predstavljene in implementirane metode so ponujene z odprtokodno licenco v okviru projekta Apertium[1]. Guat je obširneje predstavljen v razdelku 2.5.

   Prispevek je predstavljen v (Vičič, 2007a), (Vičič et al., 2009), (Vičič, 2009) in (Vičič in Homola, 2010).

[1]Apertium: machine translation toolbox, `http://sourceforge.net/projects/apertium/`

Dodatek A

Pravila prenosa

Razdelek obsega pravila prenosa iz delujočega sistema Guat. Zapisana so v formatu, ki ga uporablja Apertium in temelji na jeziku XML. Za lažje razumevanje je zapis pravil nekoliko prirejen (oznake so poslovenjene). Prikazana so pravila prenosa za jezikovni par slovenščina-srbščina, smer SL –> SR. Tako oblikovana pravila uporablja modul za strukturni prenos, ki poleg pravil uporablja še dvojezični slovar. Opis formata pravil in opisi uporabe značk so predstavljeni v Razdelku 5.2.

```
<!--prazno pravilo, uporabimo samo za prikaz-->
<rule>
  <pattern>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.1: Pravilo: prazno pravilo za samostalnik. To pravilo prebere samostalnik in ga izpiše na svoj izhod, torej ne opravi nobene spremembe.

Na Sliki A.1 je prikazano prazno pravilo, namenjeno prikazu uporabe pravil. Pravilo le prebere samostalnik in ga ponovno izpiše.

Na Sliki A.2 je prikazano pravilo ujemanja pridevnika in samostalnika v sklonu, spolu in številu. Pridevniku pripišemo iste kategorije, kot jih ima samostalnik. V komentarju je zapisan primer uporabe. Pravilo je predvsem uporabno pri napačnih

```
<!--ujemanje pridevnika in samostalnika v sklonu, spolu
in številu, samostalnik je "glavni" -->
<!--primer: rdeče drevo-->
<rule>
  <pattern>
    <pattern-item n="pridevnik"/>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="pridevnik"/>
        <clip pos="2" side="tl" part="spol"/>
        <clip pos="2" side="tl" part="število"/>
        <clip pos="2" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="stopnja"/>
        <clip pos="2" side="tl" part="določnost"/>
      </lu>
      <b pos="1"/>
        <lu>
          <clip pos="2" side="tl" part="lema"/>
          <clip pos="2" side="tl" part="samostalnik"/>
          <clip pos="2" side="tl" part="spol"/>
          <clip pos="2" side="tl" part="število"/>
          <clip pos="2" side="tl" part="sklon"/>
        </lu>
    </out>
  </action>
</rule>
```

Slika A.2: Pravilo: ujemanje pridevnika in samostalnika v sklonu, spolu in številu; pridevniku pripišemo iste kategorije, kot jih ima samostalnik.

izbirah modula za razdvoumljanje. Pri sistemih z več možnimi kandidati za prevod olajšamo delo sistema za izbiro najboljšega prevoda (Ranker).

Na Sliki A.3 je prikazano pravilo ujemanja dveh pridevnikov in samostalnika v sklonu, spolu in številu. Pridevnikoma pripišemo iste kategorije, kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

Na Sliki A.4 je prikazano pravilo ujemanja zaimka in samostalnika v sklonu, spolu in številu; zaimku pripišemo iste kategorije, kot jih ima samostalnik. Primer uporablja makro *f_concord2*, ki poskrbi za ujemanje spola in števila dveh leksikal-

```
<!--ujemanje dveh pridevnikov s samostalnikom v sklonu, spolu
in številu, samostalnik je "glavni" -->
<!--primer: mlado rdeče drevo-->
<rule>
  <pattern>
    <pattern-item n="pridevnik"/>
    <pattern-item n="pridevnik"/>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="pridevnik"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="stopnja"/>
        <clip pos="1" side="tl" part="določnost"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="pridevnik"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="2" side="tl" part="stopnja"/>
        <clip pos="2" side="tl" part="določnost"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lema"/>
        <clip pos="3" side="tl" part="samostalnik"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.3: Pravilo: ujemanje dveh pridevnikov in samostalnika v sklonu, spolu in številu; pridevnikoma pripišemo iste kategorije, kot jih ima samostalnik.

nih enot, pomembnejša pa je druga enota. Samo pravilo poskrbi še za ujemanje sklona. V komentarju je zapisan primer uporabe.

```
<!--ujemanje zaimka in samostalnika -->
<!--primer: moj avto-->
<rule>
  <pattern>
    <pattern-item n="zaimek"/>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <call-macro n="f_concord2">
      <with-param pos="2"/>
      <with-param pos="1"/>
    </call-macro>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="zaimek"/>
        <clip pos="2" side="tl" part="spol"/>
        <clip pos="2" side="tl" part="število"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="samostalnik"/>
        <clip pos="2" side="tl" part="spol"/>
        <clip pos="2" side="tl" part="število"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.4: Pravilo: ujemanje zaimka in samostalnika v sklonu, spolu in številu; zaimku pripišemo iste kategorije, kot jih ima samostalnik. Primer uporablja makro *f_concord2*.

Na Sliki A.5 je prikazano pravilo ujemanja zaimka, pridevnika in samostalnika v sklonu, spolu in številu; zaimku in pridevniku pripišemo iste kategorije, kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

Na Sliki A.6 je prikazano pravilo za prenos delov povedi, ki tvorijo prihodnjik. V slovenščini tvorimo prihodnjik s pomožnim glagolom *biti* v prihodnjiku in deležnikom na *-l*; v srbščini pa s pomožnim glagolom *hteti* in nedoločno obliko glagola. Sistem zaradi poenostavitve namesto deležnika na *-l* označi glagol v slovenščini kot

```
<!--ujemanje samostalnika in navadnega glagola v spolu
in številu -->
<!--primer avto je vozil -->
<rule>
  <pattern>
    <pattern-item n="samostalnik"/>
    <pattern-item n="vbser"/>
    <pattern-item n="glavni glagol"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="whole"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lema"/>
        <clip pos="3" side="tl" part="glavni glagol"/>
        <clip pos="3" side="tl" part="čas"/>
        <clip pos="1" side="tl" part="spol"/>
        <clip pos="1" side="tl" part="število"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.5: Pravilo: ujemanje samostalnika in navadnega glagola v spolu in številu; glagolu pripišemo iste kategorije, kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

navaden glagol v pretekliku, pravilo pa poišče poljubno obliko glagola in jo spremeni v nedoločnik. Lema pomožnega glagola *biti* se prevede v lemo pomožnega glagola *hteti*, čas pa se spremeni v sedanjik. V komentarju je zapisan primer uporabe.

Na Sliki A.7 je prikazano pravilo za prenos delov povedi, ki tvorijo prihodnjik. V slovenščini tvorimo prihodnjik s pomožnim glagolom *biti* v prihodnjiku in deležnikom na *-l*; v srbščini pa s pomožnim glagolom *hteti* in nedoločno obliko glagola. Pravilo ima samo zamenjan vrstni red leksikalnih enot s pravilom na Sliki A.6. Sis-

```
   <!-- prihodnjik 1-->
   <!-- primer: kupil bom -> kupiti ću ali kupiću-->
<rule>
  <pattern>
    <pattern-item n="glavni glagol"/>
    <pattern-item n="pomožni glagol v prihodnjiku"/>
  </pattern>
  <action>
      <let>
        <clip pos="2" side="tl" part="lema"/>
        <lit v="hteti"/>
      </let>
      <let>
        <clip pos="2" side="tl" part="čas"/>
        <lit-tag v="pres"/>
      </let>
      <let>
        <clip pos="1" side="tl" part="čas"/>
        <lit-tag v="inf"/>
      </let>

      <out>
        <lu>
          <clip pos="1" side="tl" part="lema"/>
          <clip pos="1" side="tl" part="glavni glagol"/>
          <clip pos="1" side="tl" part="čas"/>
        </lu>
        <b pos="1"/>
        <lu>
          <clip pos="2" side="tl" part="lema"/>
          <clip pos="2" side="tl" part="pomožni glagol"/>
          <clip pos="2" side="tl" part="čas"/>
          <clip pos="2" side="tl" part="oseba"/>
          <clip pos="2" side="tl" part="število"/>
        </lu>
    </out>
  </action>
</rule>
```

Slika A.6: Pravilo: pravilo za prenos delov povedi, ki tvorijo prihodnjik. Pomožni glagol sledi glavnemu.

tem zaradi poenostavitve namesto deležnika na *-l* označi glagol v slovenščini kot navaden glagol v pretekliku, pravilo pa poišče poljubno obliko glagola in jo spremeni v nedoločnik. Lema pomožnega glagola *biti* se prevede v lemo pomožnega glagola *hteti*, čas pa se spremeni v sedanjik. V komentarju je zapisan primer uporabe.

Na Sliki A.8 je prikazano pravilo za prenos nedoločnika. Nedoločnik iz slovenščine v srbščino prevedemo kot členek *da*, ki mu sledi glagol v sedanjiku. Pravilo prevede vzorec navadnega glagola v poljubni obliki, ki mu sledi glagol v nedoločniku, v členek *da*, ki mu sledi glagol v istem času, kot je prvi glagol iz izvornega dela – slovenščine. V komentarju je zapisan primer uporabe.

Na Sliki A.9 je prikazano pravilo za prenos delov povedi, ki so sestavljene iz refleksivnega glagola v prihodnjiku (sistem se naslanja na povratni svojilni zaimek). V slovenščini tvorimo prihodnjik s pomožnim glagolom *biti* v prihodnjiku in deležnikom na *-l*; v srbščini pa s pomožnim glagolom *hteti* in nedoločno obliko glagola.

Sistem zaradi poenostavitve namesto deležnika na *-l* označi glagol v slovenščini kot navadni glagol v pretekliku, pravilo pa poišče poljubno obliko glagola in jo spremeni v nedoločnik. Lema pomožnega glagola *biti* se prevede v lemo pomožnega glagola *hteti*, čas pa se spremeni v sedanjik. Svojilni zaimek se ne spreminja. V komentarju je zapisan primer uporabe.

```
   <!-- prihodnjik 1-->
   <!-- primer: bom kupil -> ću kupiti ali kupiću-->
<rule>
  <pattern>
    <pattern-item n="pomožni glagol v prihodnjiku"/>
    <pattern-item n="glavni glagol"/>
  </pattern>
  <action>
      <let>
        <clip pos="1" side="tl" part="lema"/>
        <lit v="hteti"/>
      </let>
      <let>
        <clip pos="1" side="tl" part="čas"/>
        <lit-tag v="pres"/>
      </let>
      <let>
        <clip pos="2" side="tl" part="čas"/>
        <lit-tag v="inf"/>
      </let>

      <out>
        <lu>
          <clip pos="1" side="tl" part="lema"/>
          <clip pos="2" side="tl" part="pomožni glagol"/>
          <clip pos="1" side="tl" part="čas"/>
          <clip pos="1" side="tl" part="oseba"/>
          <clip pos="1" side="tl" part="število"/>
        </lu>
        <b pos="1"/>
        <lu>
          <clip pos="2" side="tl" part="lema"/>
          <clip pos="2" side="tl" part="glavni glagol"/>
          <clip pos="2" side="tl" part="čas"/>
        </lu>
    </out>
  </action>
</rule>
```

Slika A.7: Pravilo: pravilo za prenos delov povedi, ki tvorijo prihodnjik. Pomožni glagol je pred glavnemim.

```
  <!-- Namesto infinitiva v srb = da + sedanjik -->
  <!-- primer: On želi delati (slo) -> On želi da radi (srb)  -->
<rule>
  <pattern>
    <pattern-item n="glavni glagol"/>
    <pattern-item n="glavni glagol v nedoločni obliki"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
      <b pos="1"/>
      <lu>
        <lit v="da"/>
        <lit-tag v="part"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="glavni glagol"/>
        <clip pos="1" side="tl" part="čas"/>
        <clip pos="1" side="tl" part="oseba"/>
        <clip pos="1" side="tl" part="število"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.8: Pravilo: nedoločnik.

Na Sliki A.10 je prikazano pravilo ujemanja zaimka, pridevnika in samostalnika v sklonu, spolu in številu; zaimku in pridevniku pripišemo iste kategorije, kot jih ima samostalnik. Primer uporablja makro *f_concord3*, ki poskrbi za ujemanje spola in števila treh leksikalnih enot; pomembnejša je prva enota, sledi druga. Samo pravilo poskrbi še za ujemanje sklona. V komentarju je zapisan primer uporabe.

```
<!-- prihodnjik 2 -->
<!-- primer: bom si kupil -> kupiti ću si ali kupiću si-->
<rule>
  <pattern>
    <pattern-item n="pomožni glagol v prihodnjiku"/>
    <pattern-item n="zaimek"/>
    <pattern-item n="glavni glagol"/>
  </pattern>
  <action>
    <let>
      <clip pos="1" side="tl" part="lema"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="čas"/>
      <lit-tag v="pres"/>
    </let>
    <let>
      <clip pos="3" side="tl" part="čas"/>
      <lit-tag v="inf"/>
    </let>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="pomožni glagol"/>
        <clip pos="1" side="tl" part="čas"/>
        <clip pos="1" side="tl" part="oseba"/>
        <clip pos="1" side="tl" part="število"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="3" side="tl" part="lema"/>
        <clip pos="3" side="tl" part="glavni glagol"/>
        <clip pos="3" side="tl" part="čas"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.9: Pravilo: druga oblika prihodnjika.

```
<!--ujemanje zaimka, pridevnika in samostalnika (pomemben
zaimek in samostalnik) -->
<!--primer: moj lepi avto-->
<rule>
  <pattern>
    <pattern-item n="zaimek"/>
    <pattern-item n="pridevnik"/>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <call-macro n="f_concord3">
      <with-param pos="1"/>
      <with-param pos="2"/>
      <with-param pos="3"/>
    </call-macro>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="zaimek"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lema"/>
        <clip pos="2" side="tl" part="pridevnik"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="2" side="tl" part="stopnja"/>
        <clip pos="2" side="tl" part="določnost"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lema"/>
        <clip pos="3" side="tl" part="samostalnik"/>
        <clip pos="3" side="tl" part="spol"/>
        <clip pos="3" side="tl" part="število"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.10: Pravilo: ujemanje zaimka, pridevnika in samostalnika v sklonu, spolu in številu; zaimku in pridevniku pripišemo iste kategorije, kot jih ima samostalnik.

Dodatek B

Primeri prevodov

B.1 Dobri prevodi

Primeri B.1 do B.3 kažejo dobre prevode, ki ne potrebujejo dodatnih komentarjev. Najprej je pri vsakem primeru zapisan prevod, sledi izvorno besedilo.

(B.1) *Danas je lepo vreme.*

Danes je lepo vreme.

(B.2) *Kupiti ću lep novi automobil i otići ću na more.*

Kupil bom lep nov avtomobil in odšel bom na morje.

(B.3) *Juče sam video crnu lepoticu odevenu u novi mantil.*

Včeraj sem videl črno lepotico, oblečena je bila v krznen plašč.

(B.4) *Nekad je živela devojčica.*

Nekoč je živela deklica.

(B.5) *Nekog dana joj je majka rekla:*

Nekega dne ji je mama dejala:

(B.6) *I tako je crvenkapa otišla da poseti staru majku, koja je živela u kući usred šume.*

In tako je rdeča kapica odšla obiskat staro mamo, ki je živela v hiši sredi gozda.

(B.7) *Sutra ću kupiti lep crveni automobil.*

Jutri bom kupil lep rdeč avtomobil.

B.2 Napake

Primer B.8 kaže napačno razdvoumljanje, sklon pri lemi *pištola* je bil napačno izbran: orodnik namesto tožilnika. Modul, ki skrbi za ujemanje oblikoskladenjskih kategorij, je tako pripadajočemu pridevniku pripisal napačen sklon.

(B.8) *Sutra ću kupiti veoma lepim pištoljem.*

Jutri bom kupil zelo lepo pištolo.

Primer B.9 kaže napako v dvojezičnem slovarju. Statistični model, ki je bil uporabljen pri izdelavi dvojezičnega slovarja, je napačno povezal izvorno lemo *gospod* s ciljno lemo *Čerigton* (v angleščini in tudi v slovenščini Kerrington). Obe lemi se pogosto pojavljata skupaj v oblikah „gospod Kerrington".

(B.9) *Čerington, sutra biće lep dan.*

Gospod, jutri bo lep dan.

Primer B.10 kaže napako pri prevodu dela *„bo v nedeljo gostila"*. Pravilo na Sliki A.7 sicer ponazarja prevajanje prihodnjika, vendar ga sistem v tem primeru ni uporabil, saj vzorec ni popolnoma identičen vzorcu pravila. Med pomožni in glavni glagol sta vrinjeni še besedi *„v nedeljo"*. Pravilo moramo ročno popraviti tako, da bo zajemalo tudi takšne, splošnejše vzorce.

(B.10) *Ameriška Laguna Seca biće u nedelju gostila dirku klase motoGP za VN SAD.*

Ameriška Laguna Seca bo v nedeljo gostila dirko razreda motoGP za VN ZDA.

Primer B.11 kaže napako pri prevodu *„tokratne”*; sistem je ta del prevedel v *„tokratene”*. Prevod je narejen z metodo povečevanja dvojezičnega korpusa, ki je opisana v Razdelku 4.3.2.2. Lema je samo prenesena v ciljni jezik, izbrana ji je najprimernejša paradigma. Tokrat pa obstaja prevod, ki ni podobnica in metoda ni bila uspešna.

(B.11) *Od tokratene preizkušnje mnogo očekuje i Ducatijev dvojec, jer si tako Hayden kao Rossi žele vidljiv plasman.*

Od tokratne preizkušnje veliko pričakuje tudi Ducatijev dvojec, saj si tako Hayden kot Rossi želita vidnejše uvrstitve.

Primer B.12 kaže napako pri ujemanju samostalnika in pridevnika. To napako je povzročilo ločilo, ki loči ti dve besedi in ga pravilo ne upošteva. Primer je le delno spremenjen primer B.3.

(B.12) *Juče sam vidim crnu lepoticu, odevenu je bila u krznen mantil.*

Včeraj sem videl črno lepotico, oblečena je bila v krznen plašč.

Primer B.13 kaže popolnoma zgrešen prevod: veznik *pa* bi moral biti izpuščen, pri prevodu leme *nabrati* je v ciljnem oblikoskladenjskem slovarju manjkala ustrezna oblika. Leme *cvetlica* oziroma besedne oblike *cvetlic* ni v izvornem oblikoskladenjskem slovarju.

(B.13) *Zašto ali joj ne nabrati još buket lepih cvetlic?*

Zakaj pa ji ne nabereš še šopek lepih cvetlic?

Primer B.14 kaže napačen prevod besede gospod *čerington*, ki je nastal s statistično metodo. Besedi gospod in Charington sta v več primerih v isti povedi. napačno je določen tudi spol pri besedi *loš* v ciljnem jeziku.

(B.14) *Čerington volk, taj misao međutim nije loš.*

Gospod volk, ta misel pa ni slaba.

Primer B.15 kaže „neroden” prevod *Med tem* v *Između tim*. V dvojezičnem prevajalnem slovarju manjkajo naslednje leme: *rožica, volk, hitro, babica*.

(B.15) *Između tim, kad je crvenkapa skupljala rožice, je volk hitro optrčao do babici.*

Med tem, ko je rdeča kapica nabirala rožice, je volk hitro stekel k babici.

Primer B.16 kaže primer napak v dvojezičnem slovarju. V dvojezičnem prevajalnem slovarju manjkajo naslednje leme: *rožica, volk, hitro, babica*.

(B.16) *Kucao je na vrata babičine koče. Začuo je krhki babičin glas.*

Potrkal je na vrata babičine koče. Zaslišal je slaboten babičin glas.

Primer B.17 kaže naslednje napake: namesto *što* bi moral sistem izbrati *koji*; manjkajoča beseda v dvojezičnem slovarju: *pečenka*; manjkajoče besede v enojezičnem izvornem slovarju: *pujsek, hišica, gozd*. Števila niso sklanjana.

(B.17) *Nekad je živeo volk, što si je izvanredno želeo da skuvalo pečenka iz tri pujskov, što su živeli u hišici usred šume.*

Nekoč je živel volk, ki si je neznansko želel pripraviti pečenko iz treh pujskov, ki so živeli v hišici sredi gozda.

Primer B.18 kaže naslednje napake: manjkajoča besedna oblika v ciljnem slovarju *svaki*, napačen prevod: *kako – samu*; manjkajoča beseda v izvornem slovarju: *pretental*; manjkajoč prevod v dvojezičnem slovarju: *lahek*.

(B.18) *Svaki dana je mislio, samu bi ih pretental, da bi ih lahek uhvatio.*

Vsak dan je premišljeval, kako bi jih pretental, da bi jih lahko ujel.

Primer B.19 kaže naslednje napake: manjkajoči besedni obliki v ciljnem slovarju *svaki, zaprepastiti*; napačna prevoda: *najnajprej, splezalo*, pri slednjem je napačen tudi sklon. Manjkajoča beseda v izvornem slovarju: *šlo – iti*.

(B.19) *Najnajprej je pokušao da splezalo kroz dimnjaka i ih zaprepastiti, nego sve samo nije šlo tako, kao si je zamislio.*

Najprej je poskusil splezati skozi dimnik in jih presenetiti, ampak vse le ni šlo tako, kot si je predstavljal.

Primer B.20 kaže naslednje napake: v dvojezičnem prevajalnem slovarju manjkajo naslednje leme: *volk, pokanje*; napačen prevod: *letneo*; napačen sklon pri besedi *dimnjaka*.

(B.20) *Kad je volk letneo kroz dimnjaka je začuo od unutra pokanje.*

Ko je volk lezel skozi dimnik je zaslišal od znotraj pokanje.

Primer B.21 kaže naslednje napake: v dvojezičnem prevajalnem slovarju manjka naslednja lema: *pujsek, hitro, lovec*.

(B.21) *Mislio je, da je za pujskih na poseti lovec, zato je hitro pobegao kući.*

Mislil je, da je pri pujskih na obisku lovec, zato je hitro zbežal domov.

Primer B.22 kaže naslednje napake: v dvojezičnem prevajalnem slovarju manjka naslednja lema: *pujsek*; napačen prevod: *pekelovima*; veznik *pa*.

(B.22) *Ali međutim su pujski samo kesten pekelovima.*

Vendar pa so pujski samo kostanj pekli.

Primer B.23 kaže naslednje napake: v dvojezičnem prevajalnem slovarju manjka naslednja lema: *volk*; napačna prevoda: *jedan, međutim*.

(B.23) *Jedan dana međutim si je volk rekao:*

Drugi dan pa si je volk rekel:

Primer B.24 kaže naslednje napake: v dvojezičnem prevajalnem slovarju manjka naslednja lema: *volk*; napačna prevoda: *Moguće, proverilo*; napačna oseba pri besedi *uspelo*.

(B.24) *Moguće bi mi međutim uspelo, ako bi ih pokušao da proverilo, da sam blag volk.*

Mogoče bi mi pa uspelo, če bi jih poskusil prepričati, da sem prijazen volk.

Primer B.25 kaže manjkajočo besedo v izvornem slovarju: *spustili*.

(B.25) *Zatim bi me spustili u kuću i ja bi ih uhvatio.*

Potem bi me spustili v hišo in jaz bi jih ujel.

Slike

Tabele

Algoritmi

Literatura

Lars Ahrenberg in Maria Holmqvist. Back to the Future? The Case for English-Swedish Direct Machine Translation. In *Proceedings of The Conference on Recent Advances in Scandinavian Machine Translation*. University of Uppsala, 2004.

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Laerty, Dan Melamed, Franz Josef Och, David Purdy, Noah A. Smith, in David Yarowsky. Statistical Machine Translation, Final Report. Technical report, JHU, 1999.

ALPAC. Languages and machines: computers in translation and linguistics. Technical report, National Academy of Sciences, National Research Council, 1966.

Kemal Altintas in Ilyas Cicekli. A Machine Translation System between a Pair of Closely Related Languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, 5. CRC Press, 2002.

Amebis. Jezikovne tehnologije, 2011. URL `http://www.amebis.si/`.

Apertium. Apertium: machine translation toolbox, 2010. URL `http://sourceforge.net/projects/apertium`.

Dough Arnold. *Computers and Translation: A Translator's Guide*. Benjamin Translation Library, 2003.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, in Robert L. Mercer. Readings in speech recognition. In Alex Waibel in Kai-Fu Lee, editors, *Readings in speech recognition*, poglavje A tree-based statistical language model for natural language speech recognition, 507–514. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

Jim Baker, Peter Brown, Lynn Carlson, Eduard Hovy, Charles Wayne, in John White. Machine translation evaluation methodology. Technical report, DARPA, 1992.

Satanjeev Banerjee in Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, 224–231. Seattle, WA, 2000.

Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, in François Yergeau. Extensible Markup Language (XML) 1.0 (Fifth Edition). Technical report, W3C, 2008.

Peter Brown, Peter Cocke, Stephen Della Pietra, Vincent Della Pietra, Fredrik Jelinek, John Lafferty, Robert Mercer, in Paul S. Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1994.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, in Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479, 1992.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, in Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19:163–311, 1993.

Andrea Burbank, Marine Carpuat, Stephen Clark, Markus Dreyer, Pamela Fox, Declan Groves, Keith Hall, Mary Hearne, I. Dan Melamed, Yihai Shen, Andy Way, Ben Wellington, in Dekai Wu. Final Report of the 2005 Language Engineering Workshop on Statistical Machine Translation by Parsing. Technical report, JHU, 2005.

Chris Callison-Burch, Miles Osborne, in Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*, 249–256. Association for Computational Linguistics, 2006.

Nicoletta Calzolari in Monica Monachini. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: a common proposal and applications to European languages. EAGLES Report, ILC-CNR, Pisa, Pisa: ILC., 1996.

Eugene Charniak. A maximum-entropy-inspired parser. In *ANLP*, 132–139. Association for Computational Linguistics, 2000.

Eugene Charniak. Statistical techniques for natural language parsing. *AI Magazine*, 18(4):33–44, 1997.

Harald Clahsen, Ingrid Sonnenstuhl, Meike Hadler, in Sonja Eisenbeiss. Morphological paradigms in language processing and language disorders. Essex Research Reports in Linguistics 34, University of Essex, University of Essex, Colchester, UK, 2000. URL `http://www.essex.ac.uk/linguistics/publications/errl/./hc_vbpar.pdf`.

Philip Clarkson in Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of EUROSPEECH '97*, 2707–2710. ISCA, 1997.

Jessica M. Coates. Creative Commons : the next generation : Creative Commons licence use five years on. *SCRIPTed*, 4(1):72–94, March 2007.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Michael Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.

Antonio M. Corbi-Bellot, Mikel L. Forcada, in Sergio Ortiz-Rojas. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of the EAMT conference*, 79–86. HITEC e.V., May 2005.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevič, Tomaž Erjavec, Heiki Jaan Kaalep, in Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL*, 315–319. Association for Computational Linguistics, 1998.

Helge Dyvik. Exploiting Structural Similarities in Machine Translation. *Computers and Humanities*, 28:225–245, 1995.

EAMT. European Association for Machine Translation, 2010. URL `http://www.eamt.org/`.

EGYPT. The EGYPT Statistical Machine Translation Toolkit, 2007. URL `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/`.

Katherine Eng, Alex Fraser, Daniel Gildea, Viren Jain, Zhen Jin, Shankar Kumar, Sanjeev Khudanpur, Franz Och, Dragomir Radev, Anoop Sarkar, Libin Shen, David Smith, in Kenji Yamada. Final report, syntax for statistical MT group. Technical report, JHU, 2003.

Tomaž Erjavec. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4. Conference on Language Resources and Evaluation, LREC'04*, 1535–1538. ELRA, 2004.

Tomaž Erjavec. The English-Slovene ACQUIS corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 06. ELRA, 2006.

Tomaž Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, in Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. ELRA.

Tomaž Erjavec in Saša Džeroski. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41, 2004.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, in Ralf Steinberger. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. *Archives of Control Sciences*, 15:529–540, 2005.

Tomaž Erjavec, Darja Fišer, Simon Krek, Špela Arhar, Nina Ledinek, Amanda Saksida, Breda Sivec, in Blaž Trebar. Oblikoskladenjske specifikacije JOS. Technical report, IJS - Institut Jožef Stefan, 2009.

Tomaž Erjavec, Darja Fišer, Simon Krek, in Nina Ledinek. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, volume 32(4). ELRA, 2010.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Mikel L. Forcada. Open-source machine translation: an opportunity for minor languages. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, 1–7. Genoa, Italy, 2006.

King Sun Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

GenPar. The GenPar Toolkit for Research on Generalized Parsing, 2010. URL http://nlp.cs.nyu.edu/GenPar/.

GNU. GNU General Public License, 2010. URL http://www.gnu.org/licenses/index%5Fhtml#GPL.

Google. The Google translator, 2008. URL http://www.google.com/translate_t.

Jan Hajič. RUSLAN: an MT System Between Closely Related Languages. In *Proceedings of the third conference of the European Chapter of the Association for Computational Linguistics*, 113–117. Association for Computational Linguistics, 1987.

Jan Hajič. Morphological tagging: data vs. dictionaries. In *Proceedings of the North American chapter of the Association for Computational Linguistics conference*, 2000.

Jan Hajič, Jan Hric, in Vladislav Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, 7–12. Association for Computational Linguistics, 2000.

Jan Hajič, Petr Homola, in Vladislav Kuboň. A simple multilingual machine translation system. In Eduard Hovy in Elliott Macklovitch, editors, *Proceedings of the MT Summit IX*, 157–164, New Orleans, USA, 2003. AMTA.

Morris Halle in Alec Marantz. *The View from Building 20*, poglavje Distributed Morphology and the Pieces of Inflection, 111–176. Cambridge, MA: MIT Press, 1993.

Petr Homola. *Syntactic analysis in machine translation*. Studies in Computational and Theoretical Linguistics. Institute of Formal and Applied Linguistics, 2010.

Petr Homola in Vladislav Kuboň. A method of hybrid MT for related languages. In *Proceedings of the IIS*, 269–278. Academic Publishing House EXIT, 2008a.

Petr Homola in Vladislav Kuboň. Improving machine translation between closely related Romance languages. In *Proceedings of EAMT*, 72 – 77. HITEC e.V., 2008b.

Petr Homola in Jernej Vičič. Combining MT Systems Effectively. In *Proceedings of the 23th International Florida-Artificial-Intelligence-Research-Society Conference (FLAIRS 2010)*, 198–203, Daytona Beach, Florida, USA, 2010. Florida AI Research Society, Florida AI Research Society.

Petr Homola, Vladislav Kuboň, in Jernej Vičič. *Recent Advances in Intelligent Information Systems*, poglavje Shallow Transfer Between Slavic Languages, 219–232. Academic publishing house EXIT, Warsaw, 2009.

John W. Hutchins in Harold L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.

Jan Rupnik and Miha Grčar and Tomaž Erjavec. Improving Morphosyntactic Tagging of Slovene Language through Meta-tagging. *Informatica (Slovenia)*, 34 (2):169–176, 2010.

Laura Janda. Inflectional morphology. In Dirk Geeraerts in Hubert Cuyckens, editors, *Handbook of Cognitive Linguistics*, 632–649. Oxford: Oxford U Press, 2007.

Slava Katz. Estimation of probabilities from sparse data for the language model. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

Philipp Koehn, Franz Josef Och, in Daniel Marcu. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics, 2003.

Philipp Koehn, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, in Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07)*, 177–180. Association for Computational Linguistics, 2007.

Andras Kornai. *Extended Finite State Models of Language*. Cambridge University Press, 1999.

Gorka Labaka, Nicholas Stroppa, Andy Way, in Kepa Sarasola. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. In *Proceedings of the Machine Translation Summit XI*, 41–48. EAMT, 2007.

John Lafferty, Daniel Sleator, in Davy Temperley. Grammatical trigrams: A probabilistic model of link grammar. In *In Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 89–97, 1992.

Richard J. Landis in Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174, 1977.

Alon Lavie in Michael J. Denkowski. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115, September 2009. ISSN 0922-6567.

LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, LDC, 2005.

Geoffrey Leech in Andrew Wilson. EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. Technical report, ILC-CNR, Pisa, 1996.

Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, 845–848, 1965.

Rochelle Lieber. *Deconstructing Morphology: Word Formation in Syntactic Theory*. University of Chicago Press, 1992.

Mitchell P. Marcus, Beatrice Santorini, in Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

I. Dan Melamed. Statistical machine translation by parsing. In *Proceedings of ACL*, 653–660. Association for Computational Linguistics, 2004a.

I. Dan Melamed. Algorithms for syntax-aware statistical machine translation. In *Proceedings of TMI*, 40–54. Baltimore, 2004b.

Vesna Mikolič, Jernej Vičič, in Jana Volk. *Jezikovni korpusi v medkulturni komunikaciji*, poglavje Namen in metode urejanja večjezičnega korpusa turističnih besedil (TURK), 65–74. Znanstveno-raziskovalno središče, Založba Annales, 2009.

Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*, 173–180, 1984.

Franz Josef Och. Challenges in Machine Translation. In *Proceedings of the ISCSLP*, 15. Springer, 2006.

Franz Josef Och in Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*, 29:19–51, 2003.

Franz Josef Och in Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 417–449, 2004.

Franz Josef. Och, Christoph Tillmann, in Hermann Ney. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, 20–28, University of Maryland, College Park, MD, USA, 1999.

George Orwell. *1984*. Secker and Warburg, London, 1949.

Kishore Papineni, Salim Roukos, Todd Ward, in Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. Technical report, IBM, 2001.

Mirko Popovič in Peter Willett. The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390, 1992.

Martin F. Porter. An algorithm for suffix stripping. *Program Journal*, 14:130–137, 1980.

Promt, 2010. URL `http://www.e-promt.com/`.

Lau Raymond, Ronald Rosenfeld, in Salim Roukos. Trigger-based language models using maximum likelihood estimation of exponential distributions. In *Proceedings of Speech and Signal Processing (ICASSP)*, 8. Institute of Electrical and Electronic Engineers, 1993.

Emmanuel Roche in Yves Schabes. *Finite-State Language Processing*. MIT Press, 1997.

Ronald Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1994.

Ronald Rosenfeld. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language*, 10:187–228, 1996.

Felipe Sanchez-Martinez in Mikel L. Forcada. Automatic induction of shallow-transfer rules for open-source machine translation. In Andy Way in Barbara Gawronska, editors, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, volume 2007:1, 181–190. Skovde University, September 2007.

Felipe Sanchez-Martinez in Mikel L. Forcada. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635, 2009.

Felipe Sanchez-Martinez in Hermann Ney. Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules. In Sampo Pyysalo Tapio Salakoski, Filip Ginter in Tapio Pahikkala, editors, *Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, 756–767. Springer-Verlag, August 2006. Copyright Springer-Verlag.

Felipe Sanchez-Martnez, Juan Antonio Perez-Ortiz, in Mikel L. Forcada. Integrating corpus-based and rule-based approaches in an open-source machine translation system. In Frank Van Eynde, Vincent Vandeghinste, in Ineke Schuurman, editors, *Proceedings of METIS-II Workshop: New Approaches to Machine Translation, a workshop at CLIN 17 - Computational Linguistics in the Netherlands*, 73–82, January 2007.

Edward Sapir. *Language: An Introduction to the Study of Speech*. Harcourt, Brace New York:, 1921.

Kevin P. Scannell. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, 103–109. Genoa, Italy, 2006.

Andrew Spencer. *Morphological Theory*. Blackwell Publishing, 1991.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, in Dániel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, 1–6, 2006.

Systran. Systran, 2010. URL `http://www.systran.co.uk/`.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, in Mikel L. Forcada. Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66, 2008. DOI: 10.1007/s10590-008-9044-3.

TEI-Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical report, TEI consortium, 2007.

Jože Toporišič. *Slovenska slovnica*. Založba Obzorja, Maribor, 2000.

Charles E. Townsend in Laura A. Janda. *Gemeinslavisch und Slavisch im Vergleich*. Verlag Otto Sagner, München, 2003.

Luis Villarejo, Mireia Farrus, Gema Ramírez, in Sergio Ortíz. A web-based translation service at the uoc based on apertium. In *Proceedings of IMCSIT*, 525–530. Institute of Electrical and Electronic Engineers, 2010.

Jernej Vičič. Rapid development of RBMT systems for related languages. In *Translating and the computer 29: proceedings of the twenty-ninth international conference on translating and the computer*, 162–1733. ASLIB, 2007a.

Jernej Vičič. Rapid development of RBMT systems for related languages, a case study on language pair Slovenian - Serbian. In *In Proceedings of the ERK*, 95–98. Založba FE-FRI, 2007b.

Jernej Vičič. Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages. In *Language technologies: proceedings of the conference*, 98–103. Institut »Jožef Stefan«, Ljubljana, 2008.

Jernej Vičič. *Jezikovni korpusi v medkulturni komunikaciji*, poglavje Metode hitre izdelave gradiv za prevajalne sisteme plitkega prenosa za visoko pregibne jezike, 133–153. Znanstveno-raziskovalno središče, Založba Annales, 2009.

Jernej Vičič. Strojno prevajanje in slovenščina. In *Proceedings of the 13th International Multiconference Information Society - IS 2010*, 47–52. Institut Jožef Stefan, Institut »Jožef Stefan«, Ljubljana, 2010.

Jernej Vičič in Andrej Brodnik. A method for statistical machine translation by parsing for less-used languages, 2006.

Jernej Vičič in Andrej Brodnik. A method for statistical machine translation by parsing for less-used languages. *Advances in Methodology and Statistics*, 1, 2008.

Jernej Vičič in Tomaž Erjavec. Vsak začetek je težak: avtomatsko učenje prevajanja slovenščine v angleščino. In *Language technologies, proceedings of the conference*, 20–27. Institut »Jožef Stefan«, Ljubljana, 2002.

Jernej Vičič in Mikel L. Forcada. Comparing greedy and optimal coverage strategies for shallow-transfer machine translation. In *Intelligent information systems XVI : proceedings of the International IIS '08 conference*, 307–316. Academic publishing house EXIT, Warsaw, 2008.

Jernej Vičič in Petr Homola. Speeding up the Implementation Process of a Shallow Transfer Machine Translation System. In *Proceedings of the 14th EAMT Conference*, 261–268, Saint Raphael, France, 2010. European Association for Machine Translation, HITEC e.V.

Jernej Vičič, Petr Homola, in Vladislav Kuboň. A method to restrict the blow-up of hypotheses of a non-disambiguated shallow machine translation system. In *Proceedings of the RANLP*, 460–464, Borovec, Bulgaria, 2009. Association for Computational Linguistics.

Stephan Vogel, Franz Josef Och, in Hermann Ney. The Statistical Translation Module in the Verbmobil System. In *Proceedings of the KONVENS conference*, 291–293. Springer, 2000.

Lloyd R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1–14, 2003.

Dekai Wu. MT model space: statistical versus compositional versus example-based machine translation. *Machine Translation*, 19(3-4):213–227, 2005.

Wolfgang U. Wurzel. Paradigmenstrukturbedingungen: Aufbau und Veranderung von Flexionsparadigmen. In *Proceedings of the 7th International Conference on Historical Linguistics*, 629–644. John Benjamins Pub Co, 1987.

Łukasz Dębowski, Jan Hajič, in Vladislav Kuboň. Testing the Limits – Adding a New Language to an MT System. *The Prague Bulletin of Mathematical Linguistics*, 78:95–101, 2002. ISSN 0032-6585.

Glosar

accuracy
točnost – stopnja ustreznosti merjene ali izračunane količine glede na njeno dejansko vrednost. 11

action
ukrep, ki ga izvede pravilo. 71, 76

alignment template – AT
predloga poravnave. 74

almost perfect agreement
skoraj popolno soglasje. 108

Apertium
ogrodje za postavitev sistemov za strojno prevajanje na osnovi pravil plitke analize in prenosa. 6

cognates
podobnice. 24

concatenated
združeni (morfemi). 56

concrete syntax tree
drevo izpeljav. 13

confidence score
stopnja zaupanja. 91

dummy rule
prazno pravilo. 80

fair agreement
dokajšnje ujemanje. 108

false friends
lažni prijatelji. 24

fidelity, accuracy
informativnost – določa, v kolikšni meri prevod posreduje enake informacije kot izvirnik. 95

fluency
slovnična pravilnost. 96

fully automatic machine translation – FAMT
strojno prevajanje naravnih jezikov brez uporabnikovega sodelovanja. 33

generalized parsing
posplošeno razčlenjevanje. 39

hidden markov model – HMM
skriti markovski model. 41

inflectional morphology
pregibno oblikoslovje. 9, 13

intelligibility
razumljivost – določa, ali je prevod jasen. 95

inter-rater agreement
ujemanje med ocenjevalci. 107

interlingua
vmesni jezik. 69

language style
ustreznost jezika. 96

left-to rigth longest match – LRLM
algoritem najdaljšega možnega ujemanja iz leve proti desni. 32

less used languages
manj uporabljeni jeziki. 89

long-distance dependencies
oddaljene odvisnosti. 17

mappings
povezave. 98

maximum entropy language model
model največje entropije. 16

maximum likelyhood estimation – MLE
najvišja ocena verjetnosti. 16

moderate agreement
zadovoljivo ujemanje. 108

morphosyntactic analysis
oblikoskladenjska analiza. 38

multiset
množica z dvojniki. 31

n-best set
najboljši n kandidati za prevode. 48

n-gram language model
modeli, temelječi na n-gramih, n-gramski modeli. 16

native speaker
govorec maternega jezika. 105, 107

natural language processing
statistična obdelava naravnih jezikov. 15

NIST machine translation evaluation workshop
delavnica o vrednotenju strojnega prevajanja. 100

no agreement
ni ujemanja. 108

non-native language
nematerni jezik. 100

non-terminals
neterminalnimi simboli slovnice. 13

out of domain error
napaka manjkajočih besed izven domene. 104

parse tree
drevo izpeljav. 13

pattern
vzorec, ki ga pravilo išče v besedilu. 71

perplexity
nivo začudenja. 17

post-editing
končno urejanje – dodatno delo, ki ga je treba vložiti za izdelavo dovolj dobrih prevodov. 18, 96, 101

precision
preciznost. 97

shallow structural transfer
plitki strukturni prenos. 15

skipped n-gram language model
jezikovni modeli s preskakovanjem n-gramov. 17

slight agreement
rahlo ujemanje. 108

sparse data problem
problem redkih podatkov. 15, 61

statistical language model – SLM
statistični modeli jezika. 16

statistical language modelling
statistično modeliranje jezika. 16

statistical machine translation by parsing – SMTbyP
statistično strojno prevajanje na osnovi dreves izpeljav. 35, 89

statistical parsing models
statistični modeli razčlenjevanja besedila. 89

stemming
krnjenje. 12, 98

structural transfer module
Apertiumov modul strukturnega prenosa. 50

substantial agreement
znatno ujemanje. 108

text chunks
deli besedila (pogosto tudi opisovani kot fraze, vendar brez semantične povezave). 74

translation adequacy
vsebinska ustreznost prevodov. 96

treebank
skladenjsko označeni dvojezični poravnani korpus (drevesnica). 89, 90

trigger models
jezikovni modeli na osnovi sprožil. 17

true positives
pravilno klasificirani elementi. 97

unigram precision and recall
preciznost in priklic unigramov. 98

UNIX pipe
cevi UNIX – niz procesov, ki so povezani na podlagi standardnih tokov, tako da je izhod (stdout) vsakega procesa direktni vhod (stdin) naslednjega procesa. 78

verb valency
glagolska vezava. 22

weighed finite state transducers – WFST
uteženi končni avtomati z izhodom. 41

weighted Levenshtein edit-distance
utežena Levenshteinova razdalja. 99, 109

word alignment
besedna povezava. 75

word class
besedni razred. 15

word error rate – WER
stopnja napačnih besed. 99

word recognition rate – WRR
stopnja prepoznanih besed. 100

Stvarno kazalo